

ACCB4002 A2

by Sian Kiely

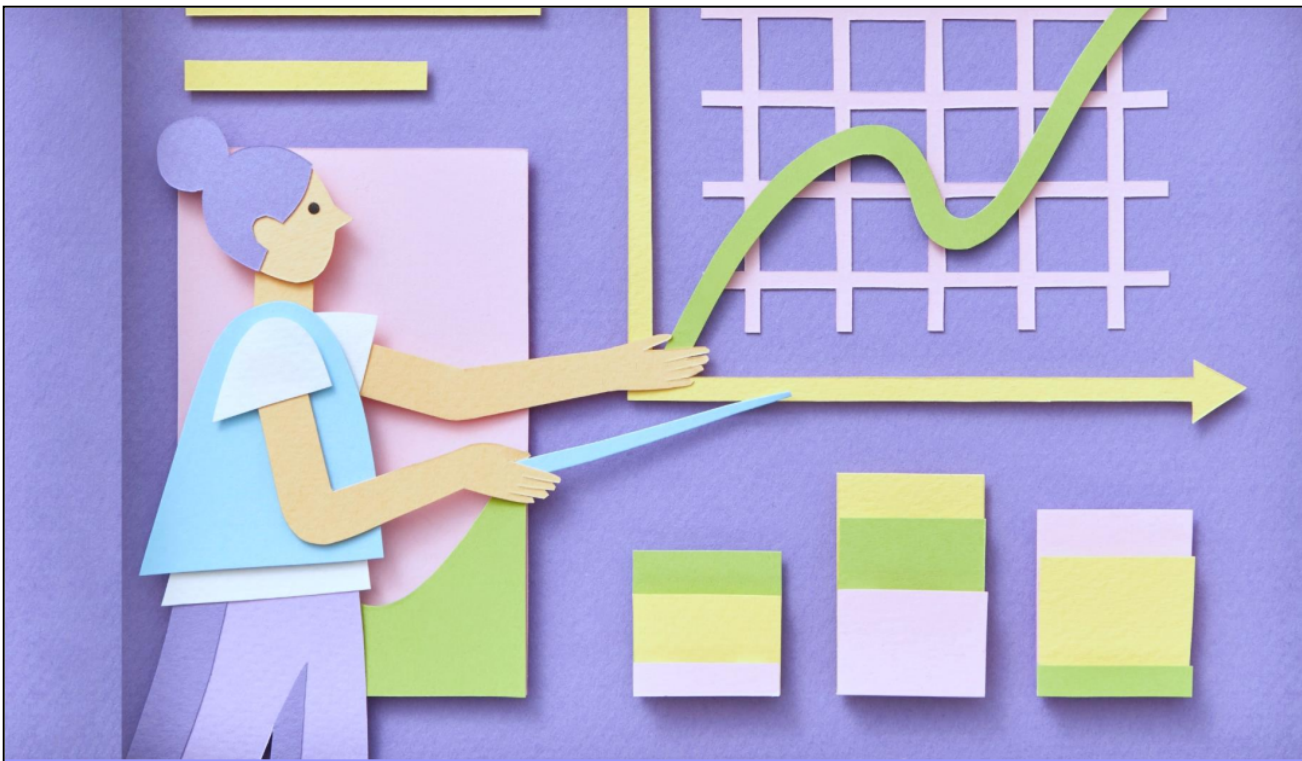
Submission date: 19-Jan-2024 02:59PM (UTC+0000)

Submission ID: 222652252

File name: 41768_Sian_Kiely_ACCB4002_A2_547400_486587022.docx (13.07M)

Word count: 2865

Character count: 15051



Statistical Analysis of Retail Store Performance

Data Analysis and Visualisation ACCB4002

Siân Eleri Kiely 2218899

2218899@student.ac.uk

Table of Contents

Introduction.....	3
Q1:	3
a) Mean, Median, Mode: Daily Sales Each Store.....	3
b) Range, Variance, Standard Deviation	4
c) Potential Outliers in Daily Sales.....	5
Q2:	6
a) 75 th percentile – sales performance upper quartile.....	6
b) Stores performing better than 75 th percentile.....	6
c) Implications of these findings for sales strategy	6
Q3:	8
a) Correlation coefficient.....	8
b) Interpret the correlation coefficient values, strength direction of relationship.....	9
c) recommendations based on findings improve customer ratings & sales Performance.	9
Q4:	10
a) Null and alternative hypotheses	10
b) t-test: significance of difference. T-statistic, degrees of freedom, p-value	10
c) Discuss implications of findings – suggest possible actions for the company.	10
Q5:	11
a) Stacked bar chart visualizing distribution of product categories.....	11
b) Scatter plot for average customer ratings on x-axis – daily sales on y-axis. Differentiate data points by product category.	11
Conclusion	11
References.....	Error! Bookmark not defined.
Table of Content.....	14
Addendum.....	15
Data Set	15

Introduction

The following report seeks to analyze retail store data, including sales, to gain insight into performance across the stores to enable business decisions. The company operates in many cities, and valuable insights may be gained from the sales data. The following table uses store 1 daily sales as an example. 12,15,18,20,20,25,28.

Q1:

Mean	Median	Mode
$\bar{x} = \frac{\Sigma x}{N}$ $\frac{12+15+18+20+20+25+28}{7} = 19.71428571428571$	12 15 18 20 20 25 28	12 15 18 20 20 25 28
$\bar{x}$ is the sample mean. [1] Σx represents the sum of all values in the sample. n is the number of data points in the sample. 2 Median, if in order, is the central most number. Mode is the number which appears most often		

See Adendum for full data set

The dataset being used in the following was obtained from five stores, looking at daily sales in **thousands of pounds per store**. John H. Kranzler [2, p.43] explains “the three main measures of central tendency are the mean, median, and mode.” Whilst mean median and mode represent average, average is usually defaulted to meaning mean, even Excel represents Mean as =AVERAGE as a formula.

a) Mean, Median, Mode: Daily Sales Each Store

Store 1		
Mean	Median	Mode
19.71428571	20	20

Each store has varying daily sale, measures of central tendency allow us to condense information into a single value, representational of a sample. They are useful when comparing level of performance.

Store 2		
Mean	Median	Mode
15	15	#N/A

We're able to analyse that on average store 5 has the highest daily sales. It's interesting to note that each store has a similar mean and median value.

Store 3		
Mean	Median	Mode
13	13	#N/A

Also noteworthy is that store 1 is the only store with a mode, this is because none of the other stores have a repeating value. This indicates that store 2 to 5 have a more varied daily sales as no two days have repeated.

Store 4		
Mean	Median	Mode
13.14285714	12	#N/A

Stores 3 and 4 have the lowest average daily sales and have the most similar mean and median.

Store 5		
Mean	Median	Mode
23.83333333	23.5	#N/A

Store 4 actually performed slightly worse with the day-to-day sales median of 12 indicating lower numbers overall.

Table 1a) store 1-5 daily sales mean median mode.

253 words

b) Range, Variance, Standard Deviation

Range		Sample Variance Population		Standard Deviation
H – L	20-10=10	$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n-1}$	$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$	$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$
Range is a simple measure of variability. Measured by subtracting the lowest (L) sale from the highest (H)		Σ = means "the sum of" $\bar{x}$ - sample mean μ - population mean <i>N refers to total number of observations</i>		X – the value in the data distribution

The above table uses store 2 daily sales as an example. 10,12,15,18,20 [2, p.48-49,54]

10	$(10-15)^2 = 25$	$25+9+0+9+25 = 68$	To calculate the variance, first subtract the mean from each number, square the result. Add each result together and divide by the number of values. N-1 is for a sample as it's an estimate. The standard deviation can be calculated by taking the Square Root of the variance.
12	$(12-15)^2 = 9$	p: $68 / 5 = 13.06$	
15	$(15-15)^2 = 0$	s: $68 / (5-1) = 17$	
18	$(18-15)^2 = 9$	p: $\sqrt{13.06} = 3.68$	
20	$(20-15)^2 = 25$	s: $\sqrt{17} = 4.123$	

Store 1		
Range	Variance	Standard Deviation
16	25.91836735	5.09100848

Store 2		
Range	Variance	Standard Deviation
10	13.6	3.687817783

Store 3		
Range	Variance	Standard Deviation
10	11.66666667	3.415650255

Store 4		
Range	Variance	Standard Deviation
17	33.26530612	5.767608354

Store 5		
Range	Variance	Standard Deviation
12	18.13888889	4.258977446

If we were to start by looking at the **range**, from store 2. We can calculate the range as 10 (see above.) Store 3 has the shared lowest range with lowest daily sale mean, but interestingly store 5 has the highest daily sale average, has the second lowest range in sales.

Variance is described as "the mean of the squares of deviation scores" [2, p.48-49,54] according to Krantzler. "We can tell how far each score deviates from the mean by subtracting the mean from it".

Standard Deviation: If the entire population is captured, the sum can be divided by N (total number of stores.) If a sample group is used, and not the entire population, an adjustment is made to the formula, dividing by $n - 1$ (in this case $5 - 1 = 4$). By dividing 68 by 4 we arrive at a variance of 17. The sample size for the retail store could be considered more indicative of a sample, however as we're analyzing the store performance it has been assumed it's the entire daily sales per store. 249 words

Table 2 Store 1-5 Range, Variance, Standard Deviation

"Two very different sets of data can have the same average, yet differ greatly in terms of how much the scores vary with every dataset"

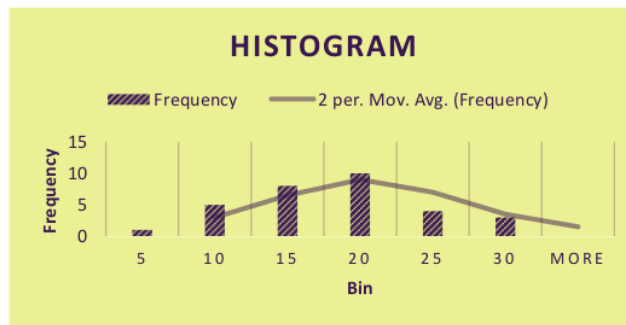
[2, p.43] John H. Krantzler

c) Potential Outliers in Daily Sales



Range C2:C13 Store 1-5 Daily Sales

Bin	Frequency
5	1
10	5
15	8
20	10
25	4
30	3
More	0



Bin Labels
<=5
6-10
11-15
16-20
21-25
26-30

Table 3 BIN & FREQUENCY

Table 4 Histogram

Table 5 Bin Labels

MIN= 5 MAX= 30

At first appearance there are no obvious outliers. To identify potential outliers, nist.gov [3] specifies it depends on the underlying distribution of the data. Frequency Distribution can be used to organise and summarise data as Kranzler [2, p.33] explains. Bins with low or no frequency could indicate an outlier. Bin breaks the data down into smaller even ranges, using this the frequency can be determined and this can be used to produce a histogram. By adding a trend line we have a frequency polygon. [2, p.40]. "The normal curve is a perfect symmetrical bell shape." Sometimes we can have skewed distributions, the above is very slightly positively skewed as we can see the greater weight on the left.

Linear regression can also be used to demonstrate each store as an independent variable in the below scatter plot. Daily sales has one potential outlier shown drifting slight from the linear trend. However no significant outliers indicate a drastic change required to the current sales strategy.

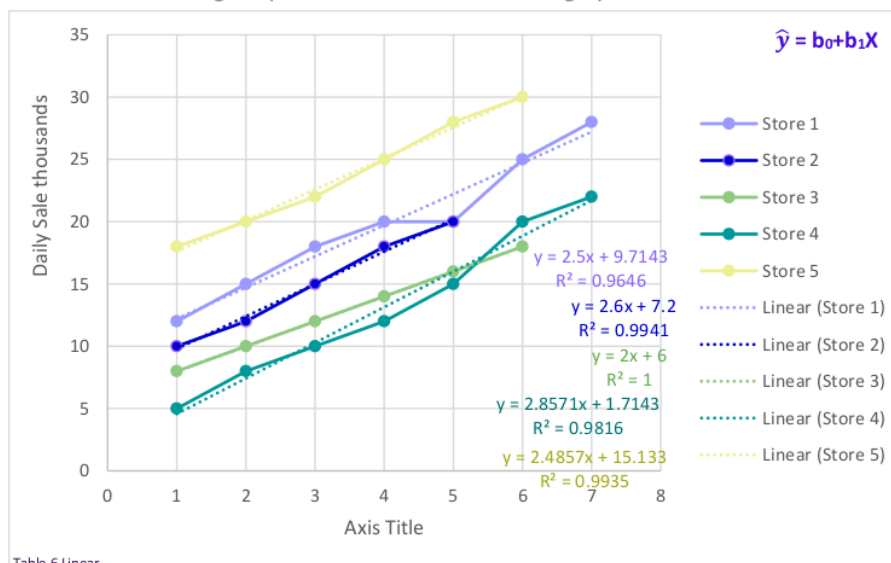


Table 6 Linear

Q2:

Store	Product Category	Daily Sales (in thousands of pounds)
1	Electronics	12
1	Electronics	15
1	Electronics	18
1	Electronics	20
1	Electronics	20
1	Electronics	25
1	Electronics	28
4	Electronics	5
4	Electronics	8
4	Electronics	10
4	Electronics	12
4	Electronics	15
4	Electronics	20
4	Electronics	22
3	Home Goods	8
3	Home Goods	10
3	Home Goods	12
3	Home Goods	14
3	Home Goods	16
3	Home Goods	18
5	Home Goods	18
5	Home Goods	20
5	Home Goods	22
5	Home Goods	25
5	Home Goods	28
5	Home Goods	30
2	Apparel	10
2	Apparel	12
2	Apparel	15
2	Apparel	18
2	Apparel	20

Table 7 75th Percentile data set

We can calculate the upper quartile, according to BBC bitesize [3] by using $\frac{3}{4}$ of the (number of value +1) or $\frac{3(n+1)}{4}$ with n representing number of data points. In this case there are 31 data points. $=3*(31+1)/4$ gives an answer of 24, and the 24th value in the ordered set is 20.

=PERCENTILE.INC(array,0.75) within excel can also be used.

a) 75th percentile – sales performance upper quartile

75th percentile of daily sales for each store				
store1	store2	store3	store4	store5
22.5	18	15.5	17.5	27.25

Store1: second highest 75th percentile more than overall upper quartile.

Store2: mid-range upper quartile

Store3: lowest overall 75th percentile value

Store4: second lowest 75th percentile value

Store5: highest 75th percentile value more than overall upper quartile.

"An individual whose score is at the 75th percentile of a group scored higher than about 75% of the persons in the norm group."
John H. Kranzler [2, p.64]

b) Stores performing better than 75th percentile.

Electronics: Store 1 performs better than the 75th percentile overall. Two thirds of the upper quartile appear against store 1.

Apparel: Store 2 is the only store selling the apparel category so no comparison can be made, but the apparel 75th percentile is lower suggesting poorer performance.

Home Goods: Store 5 is clearly performing better, home goods has the highest 75th percentile value and most of the values in the upper quartile are from store 5.

Product category 75 th percentile		
Electronics	Home Goods	Apparel
20	22.75	18

c) Implications of these findings for sales strategy

Store1 is performing better in electronic sales than store 4 and comparisons should be made to understand in how store 4 can achieve higher daily sales. Similarly, Home Goods daily sales are reaching the upper quartile more often than store 3. Store 3 is performing least favourably with a 75th percentile of 15.5. The strategy should focus on improving sales at store 3, and store 2 with Apparel having a daily sale at the 75th percentile of 18, compared with the other categories is performing the worst.

148 words



Section B

Inferential Statistics

Q3:**a) Correlation coefficient**

$$r_{xy} = \frac{\text{cov } x y}{SD_x SD_y}$$

Where $\text{cov } x y = \frac{\sum (x - \bar{x})(y - \bar{y})}{N - 1}$

A correlation Coefficient is a number between 1 and -1 describing a relationship between value pairs. Named for Karl Pearson the Pearson correlation coefficient, r , will be a value between -1.00 and +1.00 as explained by Stephen Few [5] explains. The below demonstrates arriving at a correlation coefficient value of **+0.458883366**.

$x = (x - \bar{x})$	$x = (x - \bar{x})$	xy	$(x - \bar{x})^2$	$(y - \bar{y})^2$
-4.967741935	0.161290323	-0.801248699	144	17.64
-1.967741935	0.161290323	-0.317377732	225	17.64
1.032258065	0.161290323	0.166493236	324	17.64
3.032258065	0.161290323	0.489073881	400	17.64
3.032258065	0.161290323	0.489073881	400	17.64
8.032258065	0.161290323	1.295525494	625	17.64
11.03225806	0.161290323	1.779396462	784	17.64
-6.967741935	-0.238709677	1.66326743	100	14.44
-4.967741935	-0.238709677	1.185848075	144	14.44
-1.967741935	-0.238709677	0.469719043	225	14.44
1.032258065	-0.238709677	-0.24640999	324	14.44
3.032258065	-0.238709677	-0.723829344	400	14.44
-8.967741935	0.061290323	-0.549635796	64	16.81
-6.967741935	0.061290323	-0.427055151	100	16.81
-4.967741935	0.061290323	-0.304474506	144	16.81
-2.967741935	0.061290323	-0.181893861	196	16.81
-0.967741935	0.061290323	-0.059313215	256	16.81
1.032258065	0.061290323	0.06326743	324	16.81
-11.96774194	-0.438709677	5.250364204	25	12.96
-8.967741935	-0.438709677	3.934235172	64	12.96
-6.967741935	-0.438709677	3.056815817	100	12.96
-4.967741935	-0.438709677	2.179396462	144	12.96
-1.967741935	-0.438709677	0.86326743	225	12.96
3.032258065	-0.438709677	-1.330280957	400	12.96
5.032258065	-0.438709677	-2.207700312	484	12.96
1.032258065	0.461290323	0.476170656	324	20.25
3.032258065	0.461290323	1.398751301	400	20.25
5.032258065	0.461290323	2.321331946	484	20.25
8.032258065	0.461290323	3.705202914	625	20.25
11.03225806	0.461290323	5.089073881	784	20.25
13.03225806	0.461290323	6.011654527	900	20.25

Table 8 Correlation Coefficient Calculation

$$SDx = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

$$SDy = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

N=31

=SUM($x - \bar{x}$)^2	=SUM($y - \bar{y}$)^2	=SQRT(SUM(($x - \bar{x}$)^2)/N-1)	=SQRT((SUM ($y - \bar{y}$)^2)/N-1)	$\sum xy$
10138	508.76	18.38296313	4.1180901722360	34.73870968

Cov $xy = \frac{\sum xy}{N-1}$	$\frac{34.73870968}{30-1} = 0.120603538$	=SQRT(5730.898)	R $_{xy} =$
	$18.823^2 * 4.118^2 = 5730.89875556$	75.70269979	883366

b) Interpret the correlation coefficient values, strength direction of relationship.

Using Data Analysis: correlation in Excel we can produce the following:

	12	4.2
12	1	
4.2	0.587473	1

Table 9 Correlation coefficient values

Using =CORREL(C2:C32,D2:D32) we get the result: 0.565277

A correlation coefficient value of 0.587473 indicates a moderate positive relationship. Perhaps better displayed as a scatter diagram. As Tufte writes [6], “how they differ, as the graphical display of the data makes vividly clear.” The correlation coefficient assigns a meaningful value.

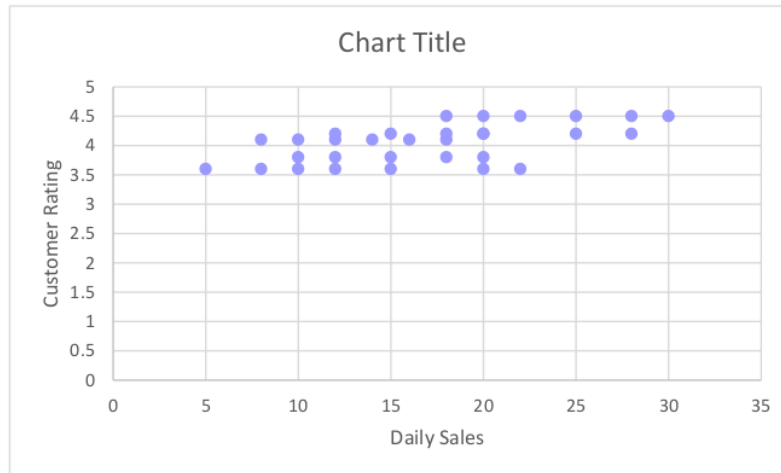


Table 10 Scatter Diagram

c) recommendations based on findings improve customer ratings & sales Performance.

The higher customer ratings are correlating with higher sales to some extent. This would suggest by improving the quality of the items for sale, daily sales would increase. Close attention should be paid to the stores performing better and the product quality and ensure lower performing stores take action to align with higher performing customer ratings and therefore results in more sales.

Q4:**a) Null and alternative hypotheses**

	sales performance rural vs urban	
null hypothesis:	there is no difference between	H0
hypothesis:	there is a significant difference in	H1

- a) In order to confirm there is a significant difference in daily sales performance between urban and rural areas. Urban and Rural will make up the two groups.

To prove the hypothesis we first need a null hypothesis. If we disprove there is no significant difference between rural and urban, we can assume the hypothesis is correct.

b) t-test: significance of difference. T-statistic, degrees of freedom, p-value

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	Variable 2
Mean	13.07692	19.77778
Variance	25.24359	33.59477
Observations	13	18
Pooled Variance	30.13911	
Hypothesized Mean Difference	0	
df	29	
t Stat	-3.35345	
P(T<=t) one-tail	0.001118	
t Critical one-tail	1.699127	
P(T<=t) two-tail	0.002235	
t Critical two-tail	2.04523	

- b) Using Excel Analysis Data, we can see

the degree of freedom is 29,

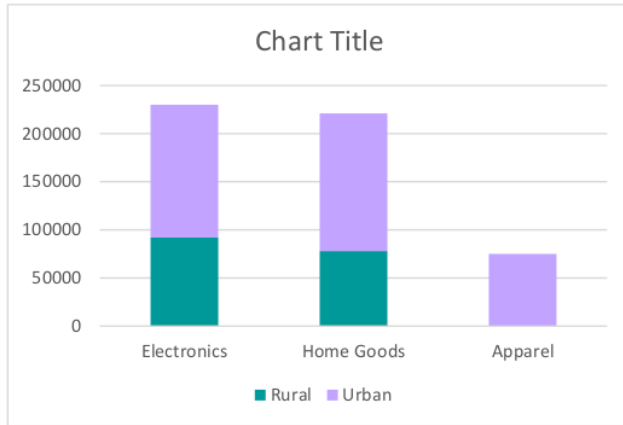
the p-values are 0.001118 and 0.002235,

the t-statistic is -3.35345.



- c) The statistical significance from the p tests are less than common significance level of 0.05, allowing us to reject the null hypothesis.

c) Discuss implications of findings – suggest possible actions for the company.

Q5:**a) Stacked bar chart visualizing distribution of product categories.****b) Scatter plot for average customer ratings on x-axis – daily sales on y-axis. Differentiate data points by product category.****Conclusion**

In conclusion the data makes it clear that store 5 is outperforming the other stores. Urban is outselling Rural against all categories. Although the variance and upper quartiles do not suggest massive discrepancies or obvious outliers. Attention should be paid to products reviewing and selling well and incentives made to improve sales at stores 2,3 and 4.

a) By creating a bar chart to visualise the distribution of product categories, we can see Apparel does not sell in rural areas. Over all Urban areas are out performing rural areas, making more daily sales in each category.

b) Using a scatter plot for average customer rating, we're able to clearly see the relationship between Daily Sales and Customer Rating.

References

- [1] Anon, "Sample Mean Formula: Definition, Applications, Examples," 1 November 2023. [Online]. Available: <https://www.pw.live/exams/school/sample-mean-formula/>. [Accessed 07 01 2024].
- [2] J. H. Kranzler, Statistics for the Terrified, 5th ed., London: Pearson Education, 2011, pp. 33, 43-49,54,64.
- [3] Anon, "1.3.5.17. Detection of Outliers," 2019. [Online]. Available: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm>. [Accessed 15 January 2024].
- [4] BBC Bitesize, "Quartiles and interquartile range - Range and box-and-whisker plots," [Online]. Available: <https://www.bbc.co.uk/bitesize/guides/zx7rk7h/revision/2>. [Accessed 10 January 2024].
- [5] S. Few, Show me the numbers, Analytics Press, 2012, p. 31.
- [6] E. R. Tufte, The visual Display of Quantitative Informaiton, 2nd ed., Connecticut, 2004, p. 14.



Table of Content

Table 1a) store 1-5 daily sales mean median mode.....	3
Table 2 Store 1-5 Range, Variance, Standard Deviation	4
Table 3 BIN & FREQUENCY	5
Table 4 Histogram	5
Table 5 Bin Labels	5
Table 6 Linear	5
Table 7 75th Percentile data set.....	6
Table 8 Correlation Coefficient Caluclation	8
Table 9 Correlation coefficient values.....	9
Table 10 Scatter Diagram	9

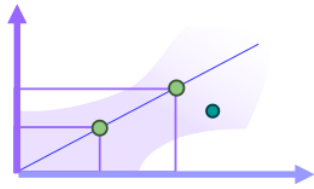
Addendum

Data Set

Store	Product Category	Daily Sales (in thousands of pounds)	Average Location Customer Rating (on a scale of 1 to 5)	Location
1	Electronics	12	4.2	Urban
1	Electronics	15	4.2	Urban
1	Electronics	18	4.2	Urban
1	Electronics	20	4.2	Urban
1	Electronics	20	4.2	Urban
1	Electronics	25	4.2	Urban
1	Electronics	28	4.2	Urban
2	Apparel	10	3.8	Urban
2	Apparel	12	3.8	Urban
2	Apparel	15	3.8	Urban
2	Apparel	18	3.8	Urban
2	Apparel	20	3.8	Urban
3	Home Goods	8	4.1	Rural
3	Home Goods	10	4.1	Rural
3	Home Goods	12	4.1	Rural
3	Home Goods	14	4.1	Rural
3	Home Goods	16	4.1	Rural
3	Home Goods	18	4.1	Rural
4	Electronics	5	3.6	Rural
4	Electronics	8	3.6	Rural
4	Electronics	10	3.6	Rural
4	Electronics	12	3.6	Rural
4	Electronics	15	3.6	Rural
4	Electronics	20	3.6	Rural
4	Electronics	22	3.6	Rural
5	Home Goods	18	4.5	Urban
5	Home Goods	20	4.5	Urban
5	Home Goods	22	4.5	Urban
5	Home Goods	25	4.5	Urban
5	Home Goods	28	4.5	Urban
5	Home Goods	30	4.5	Urban



Linear Regression example





*Lorem ipsum dolor sit amet, consectetur
adipiscing elit. Etiam aliquet eu mi quis
lacinia.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam aliquet eu mi quis lacinia. Ut fermentum a magna ut eleifend. Integer convallis suscipit ante eu varius. Morbi a purus dolor. Suspendisse sit amet ipsum finibus justo viverra blandit. Ut congue quis tortor eget sodales. Nulla a erat eget nunc hendrerit ultrices eu nec nulla. Donec viverra leo aliquet, auctor quam id, convallis orci. Sed in molestie est.

*Lorem ipsum dolor sit amet, consectetur
adipiscing elit. Etiam aliquet eu mi quis
lacinia.*

FINAL GRADE

GENERAL COMMENTS

81 /100

Question 1	20
Question 2	15
Question 3	15.5
Question 4	12
Question 5	9
Writing Skills	9

PAGE 1

PAGE 2

PAGE 3



Comment 1

Please adhere to the specified word count.



Comment 2

well analysed

PAGE 4



Comment 3

These measures help in understanding the dispersion of sales values around the mean.

PAGE 5



Comment 4

potential outliers can be found using box and whisker plot or z score method

PAGE 6



Comment 5

Q3 of 28 suggests that 75% of the daily sales in Store 5 are below 28



Comment 6

Explore strategies to boost sales on days with historically lower performance

PAGE 7

PAGE 8



Comment 7

The correlation coefficient is 0.56 which suggests a moderate positive association between average customer ratings and daily sales

PAGE 9



Comment 8

The strength and direction of the relationship need to be explicitly elucidated.



Comment 9

it can include

Customer Engagement Programs:

Loyalty programs etc..

PAGE 10



Comment 10

compare with t critical value and summarize the results



Comment 11

By adopting a flexible and adaptive strategy, the company can optimize its performance and better meet the diverse needs of customers in different locations

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17
