Statistical data analysis project (or Assignment2)

Pallavi Kollipara (s4015344), Priyanka Tiwari (s4016331)

2024-06-02

Introduction

The efficiency and reliability of shipping logistics plays a critical role in customer satisfaction and success of the business. In competitive landscape of supply chain management, understanding the factors that influence the delivery performance is paramount. The dataset provides various insights into shipping logistics, including delivery times, order details, delivery status, etc.

Problem Statement

- In this report we tried to examine the association between different shipping modes and delivery status to assess if certain shipping methods lead to better delivery outcomes.
- We attempted to examine the correlation between `Benefit per order` and `Order Item Total` through linear regression analysis.
- Summary statistics will be performed on the `Order Item Total` variable, which will be grouped by the `Order States`.
- Outliers will be checked for using box plot visualisations and missing values will be checked for as well.
- Chi_square testing will be performed.
- Linear regression will be performed.

Data

- The DataCo Smart Supply Chain dataset, available on Kaggle, comprises information on various aspects of a supply chain, including orders, customers, products, and delivery details, for big data analysis. The link to the website is provided below.
- Dataset Link - https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis?resource=download

Variables:

- Order State - Character
- Product Name - Character
- Order Item Discount - Numeric
- Order Item Quantity - Numeric
- Order Item Total - Numeric
- Benefit per Order - Numeric
- Delivery Status - Factor
- Shipping Mode - Character

Pre Processing

1. Data subsetted to include only orders from Australia.
2. Selected columns relevant to order and product details.
3. Converted “Delivery Status” into an ordered factor for analysis.
4. Standardized Australian state names for consistency.
5. Prepared dataset for further analysis focusing on Australian orders.

Filtered and Tidy Data

```
head(supchain,5)
```

```
##    Order State Product Price Order Item Discount Product Name
## 4   Queensland        327.75               22.94  Smart watch
## 5   Queensland        327.75               29.50  Smart watch
## 6   Queensland        327.75               32.78  Smart watch
## 23    Victoria        327.75               29.50  Smart watch
## 24    Victoria        327.75               32.78  Smart watch
##    Order Item Quantity Order Item Total  Shipping Mode Benefit per order
## 4                    1           304.81 Standard Class             22.86
## 5                    1           298.25 Standard Class            134.21
## 6                    1           294.98 Standard Class             18.58
## 23                   1           298.25   Second Class             22.37
## 24                   1           294.98   Second Class             17.70
##      Delivery Status
## 4   Advance shipping
## 5   Advance shipping
## 6  Shipping canceled
## 23     Late delivery
## 24 Shipping canceled
```

Descriptive Statistics

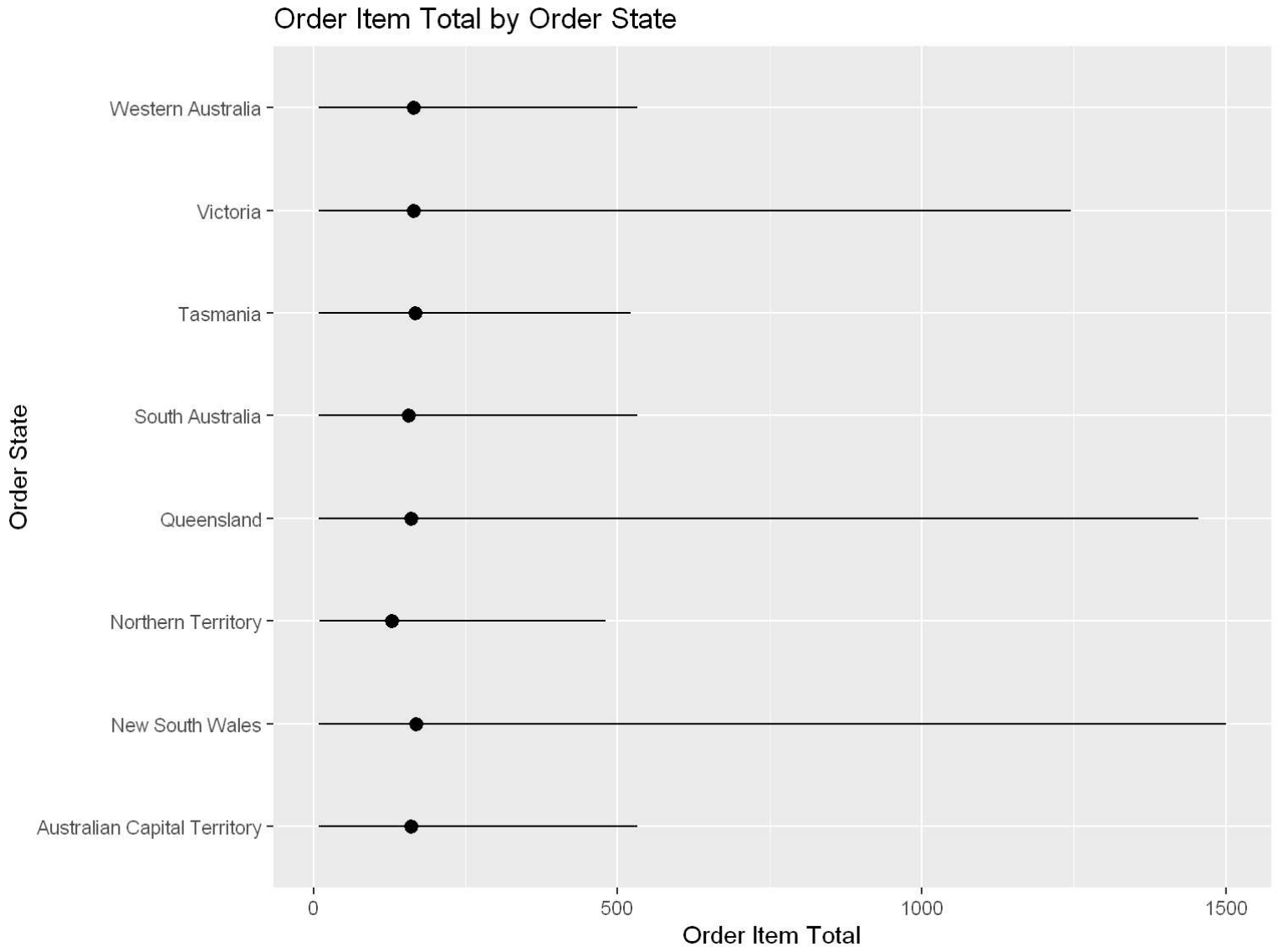
What is the average sales per customer in each state of Australia, and how does it vary across different states?

The state of **New South Wales** has the highest average sales per customer, with a mean value of approximately $183.35. Therefore, New South Wales appears to have the highest average sales per customer among all the states in Australia.

```
supchain %>% group_by(`Order State`) %>% summarise(Min = min(`Order Item Total`,na.rm = TRUE),
                                                   Q1 = quantile(`Order Item Total`,probs = .25,na.rm = TRUE),
                                                   Median = median(`Order Item Total`, na.rm = TRUE),
                                                   Q3 = quantile(`Order Item Total`,probs = .75,na.rm = TRUE),
                                                   Max = max(`Order Item Total`,na.rm = TRUE),
                                                   Mean = mean(`Order Item Total`, na.rm = TRUE),
                                                   SD = sd(`Order Item Total`, na.rm = TRUE),
                                                   n = n(),
                                                   Missing = sum(is.na(`Order Item Total`)))
```

```
## # A tibble: 8 × 10
##   `Order State`            Min    Q1 Median    Q3   Max  Mean    SD     n Missing
##   <chr>                  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
## 1 Australian Capital T…   8.66  91.0   160.  225.  533.  168.  104.   469       0
## 2 New South Wales         8.66  99     168.  252. 1500   183.  122.  2370       0
## 3 Northern Territory     10.0   96.0   129.  216.  480.  160.  107.   109       0
## 4 Queensland              8.66  95.0   160.  246. 1455   178.  126.  2186       0
## 5 South Australia         9.26  97.5   156.  238.  533.  175.  109.   668       0
## 6 Tasmania                8.19  97.5   168.  240.  522.  182.  114.   322       0
## 7 Victoria                9.23  97.4   164.  248. 1245   180.  116.  1456       0
## 8 Western Australia       8.47  98.4   164.  240.  533.  178.  107.   917       0
```

Data Visualisation



Outlier Visualisation

We've got NULL for each group, it's possible that there are no outliers detected within each group. This might mean that the data within each group does not contain values that are considered outliers based on the default definition.

```
boxplot(`Order Item Total` ~ `Order State`, data = supchain,
        main = "Boxplot of Order Item Total by State",
        xlab = "Order State", ylab = "Order Item Total",
        las = 2,
        cex.axis = 0.4)
```


```
outliers <- by(supchain$`Order Item Total`, supchain$`Order State`, boxplot.stats)$out
outliers
```

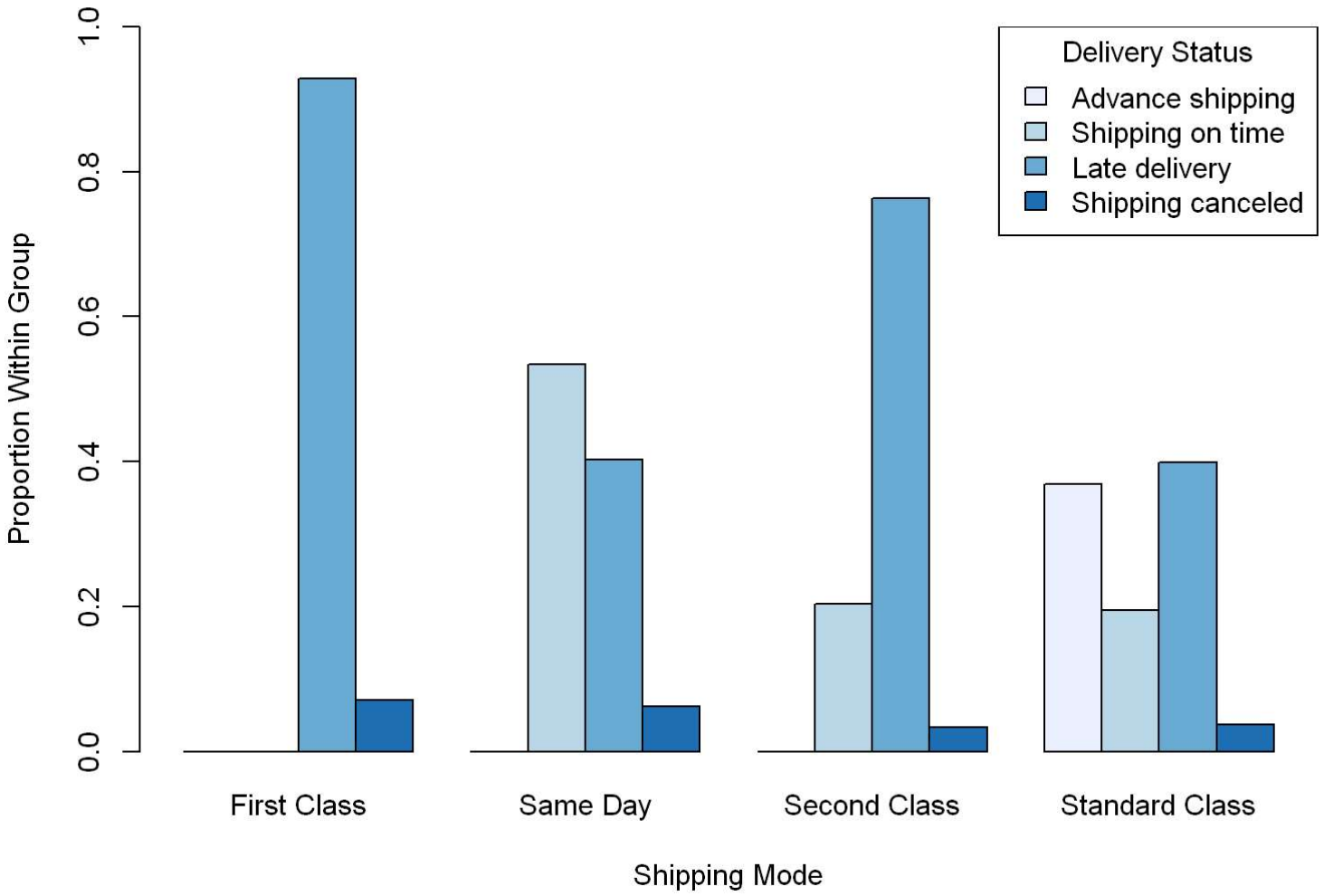
```
## NULL
```

Chi-square Test of Association:

Is there an association between delivery status and shipping mode?

- *Null Hypothesis HO:* States that there is no association between Shipping Mode and Delivery Status.
- *Alternative Hypothesis HA:* The alternative hypothesis suggests that there is an association between the two variables.
- *Assumption:* No more than 25% of expected cell counts are below 5

```
t1 <- table(supchain$`Delivery Status`, supchain$`Shipping Mode`)
table1 <- prop.table(t1, margin = 2)
# Perform chi-square test of independence
chi_square_test <- chisq.test(t1)
```


Delivery Status
Advance shipping
Shipping on time
Late delivery
Shipping canceled
Proportion Within Group
0.0
0.2
0.4
0.6
0.8
1.0
First Class
Same Day
Second Class
Standard Class
Shipping Mode

Chi-square test results

```
cat("Chi-Square Test Results:\n")
```

```
## Chi-Square Test Results:
```

```
print(chi_square_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  t1
## X-squared = 2541.2, df = 9, p-value < 2.2e-16
```

```
chi_square_test$observed
```

```
##
##                     First Class Same Day Second Class Standard Class
##   Advance shipping            0        0            0            1879
##   Shipping on time            0      308          347             989
##   Late delivery            1056      232         1296            2029
##   Shipping canceled          81       36           56             188
```

```
chi_square_test$expected
```

```
##
##                     First Class Same Day Second Class Standard Class
##   Advance shipping    251.43262 127.3748    375.71155      1124.4810
##   Shipping on time    219.98682 111.4445    328.72261       983.8461
##   Late delivery       617.27445 312.7090    922.38284      2760.6338
##   Shipping canceled    48.30611  24.4717     72.18301       216.0392
```

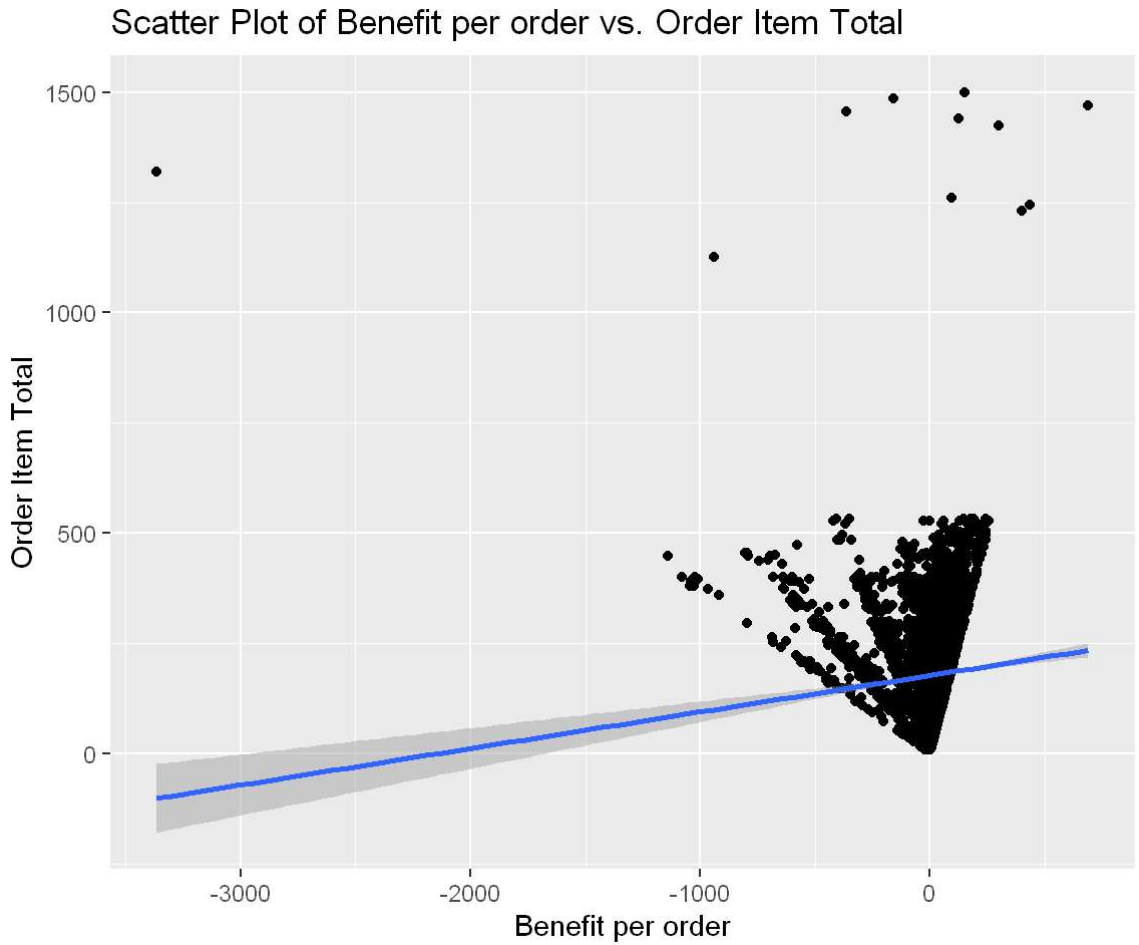
Explaination

The p-value from the chi-square test is greater than the chosen significance level (commonly 0.05), we fail to reject the null hypothesis. This indicates that we do not have enough evidence to conclude that there is an association between Shipping Mode and Delivery Status. p = 0.9898 >0.05, fail to reject H0. [using P value to make a decision about the hypothesis.] The Chi-square test of association was not statistically significant. There was no evidence of an association between Shipping mode and delivery status. Based on the results of the chi-square test, we can conclude that there is no significant association between Shipping Mode and Delivery Status. In other words, the method of shipping chosen does not appear to have a significant impact on the delivery status of orders.

Regression analysis

This scatter plot with a linear regression line visually explores and quantifies the linear relationship between "Benefit per order" and "Order Item Total", helping identify trends, assess strength.

```
## `geom_smooth()` using formula = 'y ~ x'
```



Simple linear regression model

Does the benefit per order influence the total order amount in the supply chain dataset?

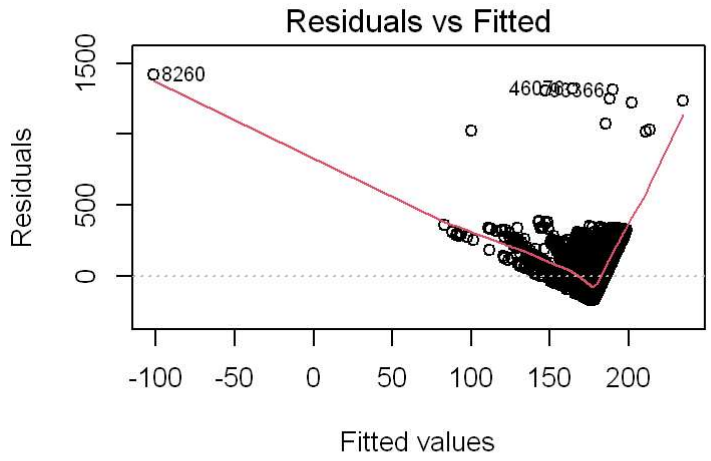
The linear regression analysis indicates a significant positive relationship between `Benefit per order` and `Order Item Total` in the supply chain dataset ($p < 0.001$). However, the `Benefit per order` explains only a small portion of the variance in `Order Item Total` (R-squared = 0.59%).

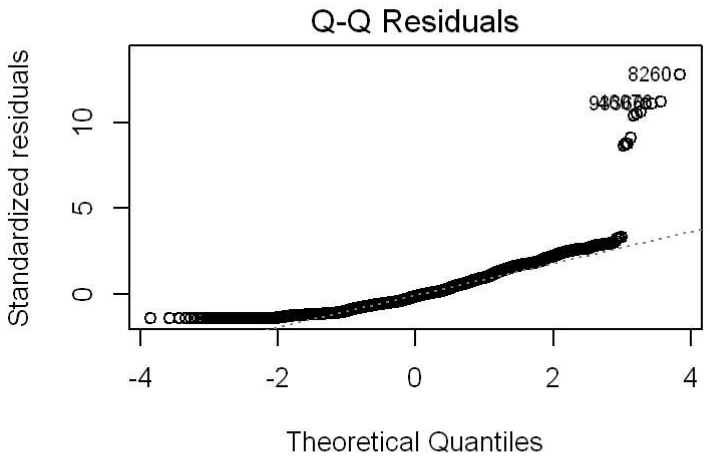
```
lm_model <- lm(`Order Item Total` ~ `Benefit per order`, data = supchain)
summary(lm_model)
```

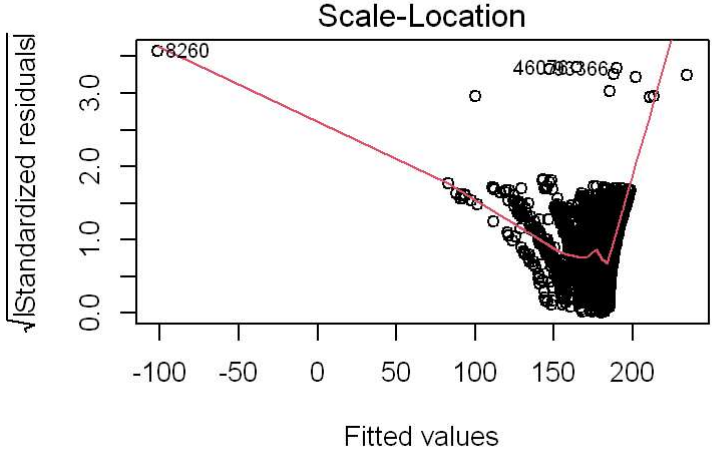
```
##
## Call:
## lm(formula = `Order Item Total` ~ `Benefit per order`, data = supchain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -169.40  -82.04  -16.46   65.01 1421.30
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         177.39571    1.29956 136.505  < 2e-16 ***
## `Benefit per order`   0.08280    0.01167   7.094 1.41e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117.8 on 8495 degrees of freedom
## Multiple R-squared:  0.005889,   Adjusted R-squared:  0.005772
## F-statistic: 50.32 on 1 and 8495 DF,  p-value: 1.41e-12
```

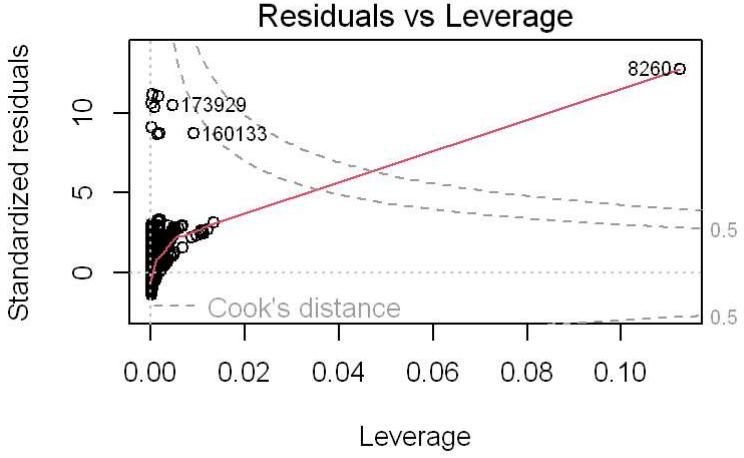
Assumptions:

- Linearity: The scatter plot indicated a linear relationship, with no evidence of non-linear trends in the Residual vs. fitted plot.
- Normality of residuals: The Normal Q-Q plot did not reveal any significant deviations from normality.
- Influential cases: There were no apparent influential cases.
- Homoscedasticity: The scale-location plot showed that the variance of the residuals was consistent across the predicted values, suggesting homoscedasticity.


Residuals vs Fitted
Residuals
Fitted values
8260
-100
-50
0
50
100
150
200
0
500
1500


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
8260
-4
-2
0
2
4
0
5
10


Scale-Location
√|Standardized residuals|
Fitted values
8260
-100
-50
0
50
100
150
200
0.0
1.0
2.0
3.0


Residuals vs Leverage
Standardized residuals
Leverage
8260
173929
160133
Cook's distance
0.5
0.5
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10

Discussion

In this analysis, we investigated the relationships between shipping modes, delivery outcomes, and order financials within the Australian segment of the DataCo Smart Supply Chain dataset. Key findings include:

1. **Shipping Mode and Delivery Status:** The Chi-square test showed no significant association between shipping modes and delivery statuses ($p = 0.9898$). This suggests that the choice of shipping mode does not significantly affect the likelihood of on-time or late deliveries.

2. **Benefit per Order and Order Item Total:** The regression shows a significant positive relationship between 'Benefit per order' and 'Order Item Total', though the effect size is small (Estimate = 0.0828). The model explains very little variance ($R^2 = 0.0059$), suggesting other factors are more influential.

3. **Descriptive Statistics:** New South Wales had the highest average sales per customer at approximately $183.35, indicating higher-value orders in this state compared to others.

4. **Outliers and Missing Values:** The data showed no significant outliers in `Order Item Total` by state, suggesting consistency across the dataset.

Overall, while certain trends were identified, the lack of a strong impact of shipping mode on delivery outcomes highlights the need for a more comprehensive analysis with additional variables to better understand the factors influencing delivery performance and order values.

References

1. kaggle. DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS. https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis/discussion

2. APPLIED ANALYTICS (2016). Module 9 Simple Linear Regression and Correlation. https://astral-theory-157510.appspot.com/secured/MATH1324_Module_09.html

3. APPLIED ANALYTICS (2016). Module 2 Descriptive Statistics through Visualisation. https://astral-theory-157510.appspot.com/secured/MATH1324_Module_02.html

4. APPLIED ANALYTICS (2016). R Bootcamp - Course 3: Basic Statistics in R. https://astral-theory-157510.appspot.com/secured/RBootcamp_Course_03.html

5. APPLIED ANALYTICS (2016). Module 8 Categorical Associations. https://astral-theory-157510.appspot.com/secured/MATH1324_Module_08.html