ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
Khoa Điện – Điện tử

BÁO CÁO BÀI TẬP LỚN
XÁC SUẤT THỐNG KÊ

Đề tài:

PHÂN TÍCH CHI PHÍ ĐẶT HÀNG Ở MỘT CỬA HÀNG ĐIỆN TỬ Ở ÚC

GVHD: Phan Thị Hường

Lớp: TN01 – Nhóm: 01

Thành phố Hồ Chí Minh – Tháng 12 năm 2024

Báo cáo phân công nhiệm vụ và kết quả thực hiện đề tài
Nhóm: 01 – Lớp: TN01 – HK: 241 – Môn: Xác suất thống kê

STT	Họ và tên	MSSV	Nhiệm vụ	Đóng góp	Điểm chia ±2
1	Phan Thiện Minh	2312107	Thống kê mô tả	100%	0
2	Nguyễn Thức Sơn	2312969	Thống kê mô tả	100%	0
3	Trương Tuấn Hải	2310887	Thống kê suy diễn	100%	0
4	Bùi Nhân Kiệt	2311747	Tổng quan dữ liệu, Kiến thức nền, Tổng hợp	100%	0
5	Cao Nguyên Phúc	2312667	Tiền xử lí số liệu	100%	0

LỜI NÓI ĐẦU

Đầu tiên, chúng em xin cảm ơn Trường Đại học Bách Khoa Thành phố Hồ Chí Minh đã tạo cơ hội cho chúng em tìm hiểu những kiến thức cơ bản của môn học Xác suất thống kê và tầm quan trọng của môn học đối với cuộc sống của chúng ta. Tiếp theo, chúng em xin gửi lời cảm ơn tới cô Phan Thị Hường, người đã dạy cho chúng em những kiến thức bổ ích và hướng dẫn chúng em vượt qua những khó khăn trong báo cáo này.

Nhóm chúng em đã có cơ hội tìm hiểu và nghiên cứu về xử lý thông tin trong công cuộc số hóa của xã hội. Và dù được cho nhiều thời gian chuẩn bị nhưng chúng em vẫn còn thiếu kinh nghiệm trong lĩnh vực này nên sai sót là không thể tránh khỏi. Vì vậy, nhóm chúng em hy vọng cô có thể giúp chúng em sửa chữa những sai sót và phản hồi để chúng em có thể tiếp tục cải thiện nghiên cứu của mình trong tương lai.

Mục lục

Danh mục hình ảnh

Danh mục các bảng

1. TỔNG QUAN DỮ LIỆU

1.1. Mô tả dữ liệu

Tập dữ liệu cung cấp trong BTL chứa thông tin về một cửa hàng điện tử trực tuyến. Cửa hàng có ba kho để giao hàng cho khách hàng. Dựa vào dữ liệu đã cho để tìm ra các mối quan hệ giữa các biến, từ đó đưa ra phỏng đoán và xây dựng mô hình dự báo chi phí đặt hàng.

- Nguồn dữ liệu lấy tại đây.
- Tiêu đề: Transactional Retail Dataset of Electronics Store.
- Thông tin dữ liệu:
  — Tác giả: SHAHRAYAR.
  — Thời gian dữ liệu được công bố: 3 năm trước.
- Giá trị quan trắc: 500 giá trị.
- Số lượng biến trong dữ liệu: 16 biến.

Trong nguồn dữ liệu gồm 2 file dữ liệu là: **dirty_data.csv** và **missing_data.csv.** Trong bài tập lớn này, chúng ta chọn file **dirty_data.csv** để làm dữ liệu cho việc xây dựng các mô hình thống kê.

1.2. Phân loại biến

- Các biến chính trong bộ dữ liệu:

Tên biến	Đơn vị	Ghi chú
delivery_charges	USD	Chi phí vận chuyển
customer_lat	-	Vĩ độ vị trí của khách hàng
customer_long	-	Kinh độ vị trí của khách hàng
order_price	USD	Giá của đơn hàng trước khi áp dụng giảm giá hoặc phí giao hàng.
season	-	Mùa mà đơn hàng được đặt
is_expedited_delivery	-	Khách hàng có hay không yêu cầu giao hàng nhanh
distance_to_nearest_warehouse	km	Khoảng cách đến kho hàng gần nhất

Bảng 1.1. Các biến chính trong bộ dữ liệu

- Mục tiêu phân tích
  — So sánh tổng chi phí đặt hàng ở các mùa.
  — Thực hiện dự báo chi phí vận chuyển các đơn đặt hàng.

2. KIẾN THỨC NỀN

2.1. Hồi quy tuyến tính bội

Mô hình hồi quy bội (Multiple Regression Model) là một công cụ mạnh mẽ trong thống kê và phân tích dữ liệu, giúp nghiên cứu mối quan hệ giữa một biến phụ thuộc và nhiều biến độc lập. Dưới đây là cơ sở lý thuyết và một số tài liệu tham khảo liên quan đến mô hình hồi quy bội. Mô hình này mở rộng từ hồi quy tuyến tính đơn giản, trong đó chỉ có một biến độc lập.

— Biến phụ thuộc (Dependent Variable): Là biến mà bạn muốn dự đoán hoặc giải

thích.

— Biến độc lập (Independent Variables): Là các biến được sử dụng để dự đoán hoặc giải thích biến phụ thuộc.

— Mô hình hồi quy tuyến tính bội có dạng tổng quát như sau:

$$\boldsymbol{Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + u}$$

* ***Trong đó:***

+ Y: Biến phụ thuộc
+ Xi: Biến độc lập
+ β1: Hệ số tự do (hệ số chặn)
+ βi: Hệ số hồi quy riêng.

Như vậy, "Hồi quy tuyến tính" là một phương pháp để dự đoán giá trị biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X).

2.1.1. Hàm hồi quy tổng thể (PRF - Population Regression Function)

Với Y là biến phụ thuộc X_2, X_3, …, X_k là biến độc lập, Y là ngẫu nhiên và có một phân phối xác suất nào đó.

→ Tồn tại E (Y | X_2, X_3, …, X_k) = giá trị xác định.

→ F (X_2, X_3, …, X_k) = E (Y | X_2, X_3, …, X_k) là hàm hồi quy tổng thể của Y theo X_2, X_3, …, X_k

Với một cá thể i, tồn tại (X2, i, X3, i, . . ., Xk, i, Yi)

— Ta có: $Y_i \neq F (X_2, X_3, \ldots, X_k) \Rightarrow u_i = Y_i - F$

— Do vậy: $Y_i = E (Y | X_2, X_3, \ldots, X_k) + u_i$

Hồi quy tổng thể PRF:

— Y = E (Y | X) + U

— E (Y | X) = F(X)

2.1.2. Hàm hồi quy mẫu (SRF - Sample Regression Function):

Do không biết tổng thể, nên chúng ta không biết giá trị trung bình tổng thể của biến phụ thuộc là đúng ở mức độ nào. Do vậy chúng ta phải dựa vào dữ liệu mẫu để ước lượng. Trên một mẫu có n cá thể, gọi $\hat{Y} \neq \hat{F}$ (X_2, X_3, …, X_k) là hồi quy mẫu. Với một cá thể mẫu $Y_i \neq \hat{F}$ ($X_{2,i}$, $X_{3,i}$, …, $X_{k,i}$) sinh ra e_i – Y_i – $\hat{F}$ (X_2, X_3, … , X_k); e_i gọi là phần dư SRF. Ta có hàm hồi quy mẫu tổng quát được viết dưới dạng như sau:

$$\boldsymbol{\hat{y}_i = \widehat{\beta_1} + \widehat{\beta_2} x_{2,i} + \widehat{\beta_3} x_{3,i} + \cdots + \widehat{\beta_k} x_{k,i}}$$

Phần dư sinh ra: $e_i = y_i - \hat{y}_i$

Ký hiệu: $\widehat{\beta_m}$ là ước lượng của β_m.

Chúng ta trông đợi $\widehat{\beta_m}$ là ước lượng không chệch của β_m, hơn nữa phải là một ước lượng hiệu quả.

Ước lượng SRF: chọn một phương pháp nào đó để ước lượng các tham số của F qua việc tìm các tham số của $\hat{F}$ và lấy giá trị quan sát của các tham số này làm giá trị xấp xỉ cho tham số của F.

2.1.3. Các giả thiết của phương pháp bình phương nhỏ nhất cho mô hình hồi quy tuyến tính bội

Trong khi xây dựng mô hình hồi quy đa biến cần kiểm tra các giả thiết như sau:

a. Hàm hồi quy là tuyến tính theo các tham số. Điều này có nghĩa là quá trình thực hành

hồi quy trên thực tế được miêu tả bởi mối quan hệ dưới dạng: y = $\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + u$ hoặc mối quan hệ thực tế có thể được viết lại ví dụ như dưới dạng lấy loga cả hai vế

b. Kỳ vọng của các yếu tố ngẫu nhiên u_i bằng 0.

Trung bình tổng thể sai số là bằng 0. Do hàm xem như là đường trung bình nên có thể giả định rằng các sai số ngẫu nhiên trên sẽ bị loại trừ nhau, ở mức trung bình, trong tổng thể.

c. Các sai số độc lập với nhau.

d. Các sai số có phương sai bằng nhau. Tất cả giá trị u được phân phối giống nhau với cùng phương sai σ^2, sao cho: Var (u_i) = E (u_i^2) = σ^2

e. Các sai số có phân phối chuẩn. Điều này rất quan trọng khi phát sinh khoảng tin cậy và thực hiện kiểm định giả thuyết trong những phạm vi mẫu là nhỏ. Nhưng phạm vi mẫu lớn hơn, điều này trở nên không mấy quan trọng.

2.1.4. Độ phù hợp của mô hình

Để có thể biết mô hình giải thích được như thế nào hay bao nhiêu % biến động của biến phụ thuộc, người ta sử dụng R^2.

Ta có:

— $\Sigma(y_i - \bar{y})^2$: TSS – Total Sum of Squares

— $\Sigma(\hat{y}_i - \bar{y})^2$: ESS – Explained Sum of Squares

— Σe_i^2: RSS – Residual Sum of Square

Ta có thể viết: TSS = ESS + RSS

Ý nghĩa của các thành phần:

— TSS là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát Y_i và giá trị trung bình.

— ESS là tổng bình phương của tất cả các sai lệch giữa các giá trị của biến phụ thuộc Y nhận được từ hàm hồi quy mẫu và giá trị trung bình của chúng. Phần này đo độ chính xác của hàm hồi quy.

— RSS là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát Y và các giá trị nhận được từ hàm hồi quy.

— TSS được chia thành 2 phần: một phần do ESS và một phần do RSS gây ra.

R^2 được xác định theo công thức:

$$\boldsymbol{R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}}$$

Tỷ số giữa tổng biến thiên được giải thích bởi mô hình cho tổng bình phương cần được giải thích được gọi là hệ số xác định, hay là trị thống kê "good of fit". Từ định nghĩa R^2chúng ta thấy R^2đo tỷ lệ hay số % của toàn bộ sai lệch Y với giá trị trung bình được giải thích bằng mô hình. Khi đó người ta sử dụng R^2để đo sự phù hợp của hàm hồi quy:

— $0 \leq R^2 \leq 1$.

— R^2cao nghĩa là mô hình ước lượng được giải thích được một mức độ cao biến động của biến phụ thuộc.

— Nếu R2 = 1, nghĩa là đường hồi quy giải thích 100% thay đổi của y.

— Nếu R2 = 0, nghĩa là mô hình không đưa ra thông tin nào về sự thay đổi của biến phụ thuộc y.

2.1.5. Khoảng tin cậy và kiểm định các hệ số hồi quy

a. Ước lượng khoảng tin cậy đối với các hệ số hồi quy:

Mục đích của phân tích hồi quy không phải chỉ suy đoán về $\beta_1, \beta_2, \dots, \beta_k$ mà còn phải kiểm tra bản chất sự phụ thuộc. Do vậy cần phải biết phân bố xác suất của β_1, β_2,…, β_k. Các phân bố này phụ thuộc vào phân bố của các u_i.

Với các giả thiết OLS, $u_i$có phân phối N$(0; \sigma^2)$. Các hệ số ước lượng tuân theo phân phối chuẩn:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} \sim \boldsymbol{N}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}\,; \boldsymbol{Se}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}\right)\right)$$

$$\frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} - \boldsymbol{\beta}_{\boldsymbol{j}}}{\boldsymbol{Se}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}\right)} \sim \boldsymbol{T}(\boldsymbol{n}-\boldsymbol{k})$$

Ước lượng phương sai sai số dựa vào các phần dư bình phương tối thiểu. Trong đó k là số hệ số có trong phương trình hồi quy đa biến:

— Ước lượng 2 phía, ta tìm được $t_{\frac{\alpha}{2}}(n-k)$ thỏa mãn:

$$\boldsymbol{P}\left(-\boldsymbol{t}_{\frac{\boldsymbol{\alpha}}{\boldsymbol{2}}}(\boldsymbol{n}-\boldsymbol{k})\right) \leq \frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} - \boldsymbol{\beta}_{\boldsymbol{j}}}{\boldsymbol{Se}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}\right)} \leq \boldsymbol{P}\left(\boldsymbol{t}_{\frac{\boldsymbol{\alpha}}{\boldsymbol{2}}}(\boldsymbol{n}-\boldsymbol{k})\right) = \boldsymbol{1}-\boldsymbol{\alpha}$$

— Khoảng tin cậy 1 − α của β_j là:

$$\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} - \boldsymbol{t}_{\frac{\boldsymbol{\alpha}}{\boldsymbol{2}}}(\boldsymbol{n}-\boldsymbol{k})\boldsymbol{Se}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}\right);\ \widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} + \boldsymbol{t}_{\frac{\boldsymbol{\alpha}}{\boldsymbol{2}}}(\boldsymbol{n}-\boldsymbol{k})\boldsymbol{Se}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}\right)\right]$$

b. Kiểm định giả thiết đối với $\boldsymbol{\beta_j}$

Kiểm định ý nghĩa thống kê của các hệ số hồi quy có ý nghĩa hay không: kiểm định rằng biến giải thích có thực sự ảnh hưởng đến biến phụ thuộc hay không. Nói cách khác là hệ số hồi quy có ý nghĩa thống kê hay không.

Có thể đưa ra giả thiết nào đó đối với β_j, chẳng hạn $\beta_j = \beta_{j^*}$. Nếu giả thiết này đúng thì:

$$\boldsymbol{T} = \frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} - \boldsymbol{\beta}_{\boldsymbol{j}}}{\boldsymbol{Se}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}\right)} \sim \boldsymbol{T}(\boldsymbol{n}-\boldsymbol{k})$$

Loại giả thiết	Giả thiết H_0	Giả thiết đối H_1	Miền bác bỏ
Hai phía	$\beta_j = \beta_{j^*}$	$\beta_j \neq \beta_{j^*}$	$\lvert t\rvert > t_{\frac{\alpha}{2}}(n-k)$
Phía phải	$\beta_j \leq \beta_{j^*}$	$\beta_j > \beta_{j^*}$	$t > t_{\alpha}(n-k)$
Phía trái	$\beta_j \geq \beta_{j^*}$	$\beta_j < \beta_{j^*}$	$t < t_{\alpha}(n-k)$

Bảng 2.1. Bảng tóm tắt giả thuyết và miền bác bỏ tương ứng

Ta có thể sử dụng giá trị P-value: P-value < mức ý nghĩa thì bác bỏ giả thiết H_0

Kiểm định β_j:

— Giả thuyết H_0: $\beta_j = 0 \Leftrightarrow x_j$ không tác động

— Giả thuyết H_1: $\beta_j \neq 0 \Leftrightarrow x_j$ có tác động.

+ βj < 0 ⇔ xj có tác động ngược
+ βj > 0 ⇔ xj có tác động thuận

2.1.6. Kiểm định mức độ ý nghĩa chung của mô hình (trường hợp đặc biệt của kiểm định WALD)

a. Khái quát về kiểm định WALD.

Giả sử chúng ta có 2 mô hình dưới đây:

(U): Y = $\boldsymbol{\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u}$

(R): Y = $\boldsymbol{\beta_1 + \beta_2 X_2 + v}$

Mô hình U được gọi là mô hình không giới hạn (Unrestrict), và mô hình R được gọi là mô hình giới hạn (Restrict). Đó là do β_3 và β_4 buộc phải bằng 0 trong mô hình R. Ta có thể kiểm định giả thuyết liên kết $\beta_3 = \beta_4 = 0$ với giả thuyết đối là ít nhất một trong những hệ số này không bằng 0.

Kiểm định giả thuyết liên kết này được gọi là kiểm định Wald, thủ tục như sau:

— Đặt các mô hình giới hạn và không giới hạn là:

(U): Y = $\boldsymbol{\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_m X_m + \beta_{m+1} X_{m+1} + \cdots + \beta_k X_k + u}$

(R): Y = $\boldsymbol{\beta_1 + \beta_2 X_2 + \cdots + \beta_m X_m + v}$

Mô hình (R) có được bằng cách bỏ bớt một số biến ở mô hình (U) đó là**:**

$$\boldsymbol{X_{m+1}\ (X_{m+1} + \cdots + X_k)}$$

— Giả thuyết H0: $\beta_{m+1} = \cdots = \beta_k = 0$

— Giả thuyết H1: "Không phải đồng thời các tham số bằng 0".

Lưu ý rằng (U) chứa k hệ số hồi quy chưa biết và (R) chứa m hệ số hồi quy chưa biết. Do đó, mô hình R có ít hơn (k − m) thông số so với U. Câu hỏi chúng ta nêu ra là (k − m) biến bị loại ra có ảnh hưởng liên kết có ý nghĩa đối với Y hay không.

Trị thống kê kiểm định đối với giả thiết này là:

$$\boldsymbol{F_C = \frac{\frac{RSS_R - RSS_U}{k-m}}{\frac{RSS_u}{n-k}} \sim F(\alpha, k-m, n-k) = \frac{\frac{{R_U}^2 - {R_R}^2}{k-m}}{\frac{1-{R_U}^2}{n-k}}}$$

Với R_2 là số đo độ thích hợp không hiệu chỉnh. Với giả thuyết không, F_C có phân phối F với (k − m) bậc tự do đối với tử số và (n − k) bậc tự do đối với mẫu số.

Bác bỏ giả thuyết H0 khi:

$$\boldsymbol{F_C > F(\alpha, k-m, n-k)}$$

Hoặc giá trị p-value của thống kê F nhỏ hơn mức ý nghĩa cho trước.

b. Kiểm định ý nghĩa của mô hình.

Trong mô hình hồi quy đa biến, giả thuyết "không" cho rằng mô hình không có ý nghĩa được hiểu là tất cả các hệ số hồi quy riêng đều bằng 0.

Ứng dụng kiểm định Wald (thường được gọi là kiểm định F) được tiến hành cụ thể như sau:

— ***Bước 1:***

+ Giả thuyết H0: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$
+ Giả thuyết H1: "có ít nhất một trong những giá trị β khác không".

— ***Bước 2:***

+ Trước tiên hồi quy Y theo một số hạng không đổi và $X_2, X_3, \ldots, X_k$, sau đó tính tổng bình phương sai số RSSU, RSSR. Phân phối F là tỷ số của hai biến ngẫu nhiên phân phối khi bình phương độc lập. Điều này cho ta trị thống kê:

$$F_C = \frac{\frac{RSS_R - RSS_U}{k-m}}{\frac{RSS_u}{n-k}} \sim F(\alpha, k-m, n-k)$$

+ Vì H_0: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$, nhận thấy rằng trị thống kê kiểm định đối với giả thuyết này sẽ là:

$$F_C = \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}} \sim F(\alpha, k-1, n-k)$$

— ***Bước 3:*** Tra số liệu trong bảng F tương ứng với bậc tự do (k − 1) cho tử số và (n − k) cho mẫu số, và với mức ý nghĩa α cho trước.

— ***Bước 4:*** Bác bỏ giả thuyết H0 ở mức ý nghĩa α nếu $F_C > F(\alpha, k-1, n-k)$

Đối với phương pháp giá trị p-value, tính giá trị p = P (F > Fc | H_0) và bác bỏ giả thuyết H_0 nếu p bé hơn mức ý nghĩa α.

2.2. Phân tích phương sai hai nhân tố có lặp

Mục tiêu của phân tích phương sai là so sánh trung bình của nhiều nhóm (tổng thể) dựa trên các giá trị trung bình của các mẫu quan sát từ các nhóm này, và thông qua kiểm đinh giả thuyết để kết luận sự bằng nhau của các trung bình tổng thể này. Trong nghiên cứu, phân tích phương sai được xem như một công cụ để xem xét sự ảnh hưởng của một yếu tố nguyên nhân (định tính) đến một yếu tố kết quả (định lượng).

Ta có các mô hình phân tích phương sai như sau: phân tích phương sai một nhân tố, 2 nhân tố và 3 nhân tố. Cụm từ nhân tố cho ta số lượng nhân tố nguyên nhân ảnh hưởng đến kết quả ta nghiên cứu.

— Phân tích phương sai hai nhân tố sử dụng để nghiên cứu tác động:

+ Hai nhân tố được quan tâm trên một biến phụ thuộc (biến giải thích).
+ Tương tác giữa các mức khác nhau của hai nhân tố.

Các giả thiết cho bài toán phân tích phương sai hai nhân tố có lặp

— Đối với nhân tố A:

+ $H_{0A}: \mu_{A1} = \mu_{A2} = \cdots = \mu_{Aa}$
+ $H_{1A}: \exists \mu_{Ai} \neq \mu_{Aj} (i \neq j)$

— Đối với nhân tố B

+ $H_{0B}: \mu_{B1} = \mu_{B2} = \cdots = \mu_{Bb}$
+ $H_{1B}: \exists \mu_{Bi} \neq \mu_{Bj} (i \neq j)$

— Đối với sự tương tác giữa A và B:

+ $H_{0AB}: Không\ có\ sự\ tương\ tác\ giữa\ A\ và\ B$
+ $H_{1AB}: Có\ sự\ tương\ tác\ giữa\ A\ và\ B$

Bảng mô hình phân tích phương sai hai nhân tố có lặp

Nguồn của sự biến thiên	SS	Df	MS	F
Giữa các nhóm	SSG	a-1	MSG	$F_A = \frac{MSG}{MSE}$
Giữa các khối	SSB	b-1	MSB	$F_B = \frac{MSB}{MSE}$
Tương tác	SSI	(a-1)(b-1)	MSI	$F_{AB} = \frac{MSI}{MSE}$
Sai số	SSE	N-ab	MSE	
Tổng cộng	SST	N-1		

Bảng 2.2. Mô hình phân tích phương sai hai nhân tố có lặp

Bác bỏ H_{0A} khi $F_A > F_{\alpha,(\alpha-1),(N-ab)}$

Bác bỏ H_{0B} khi $F_B > F_{b,(b-1),(N-ab)}$

Bác bỏ H_{0AB} khi $F_{AB} > F_{\alpha,(\alpha-1),(b-1),(N-ab)}$

3 TIỀN XỬ LÍ SỐ LIỆU

3.1 Đọc dữ liệu

Trước tiên, chúng ta cần đưa dữ liệu từ file gốc vào chương trình, từ đó mới tiến hành các bước xử lý ở lúc sau. Sử dụng lệnh read.csv để đọc dữ liệu và lưu vào biến dữ liệu, và dùng lệnh head sẽ in ra cho chúng ra số lượng dòng cần thiết. Để coi sơ lược dữ liệu, ta chọn in ra 10 dòng đầu tiên từ datasheet.

	order_id	customer_id	date	nearest_warehouse	shopping_cart
1	ORD182494	ID6197211592	6/22/2019	Thompson	[('Lucent 330S', 1), ('Thunder line', 2), ('iStream', 2), ('pearTV',...
2	ORD395518	ID0282825849	12/29/2019	Thompson	[('Thunder line', 1), ('Universe Note', 2)]
3	ORD494479	ID0579391891	3/2/2019	Nickolson	[('Thunder line', 1), ('pearTV', 2)]
4	ORD019224	ID4544561904	1/12/2019	Nickolson	[('Universe Note', 1), ('Alcon 10', 2), ('Olivia x460', 1), ('iAssist ...
5	ORD104032	ID6231506320	11/28/2019	Nickolson	[('Universe Note', 1), ('Olivia x460', 1), ('iStream', 1), ('Toshika ...
6	ORD146760	ID0311654900	9/16/2019	Bakers	[('Thunder line', 2), ('Universe Note', 1)]
7	ORD337984	ID3394768956	9/14/2019	Thompson	[('Candle Inferno', 1), ('Alcon 10', 1), ('Toshika 750', 1)]
8	ORD072312	ID0774517121	5/23/2019	Thompson	[('Universe Note', 1), ('Thunder line', 2), ('iStream', 1)]
9	ORD377837	ID4769265355	10/9/2019	Bakers	[('Alcon 10', 2), ('Thunder line', 1), ('Candle Inferno', 2), ('iAssi...
10	ORD462194	ID5301568579	3/21/2019	Thompson	[('Universe Note', 1), ('Lucent 330S', 1), ('Toshika 750', 2)]

Hình 3.1. Kết quả 10 dòng đầu tiên của dirty_data

Chúng ta có thể kiểm tra cấu trúc của tệp tin bằng lệnh str.

```
'data.frame':   500 obs. of  16 variables:
 $ order_id                    : chr  "ORD182494" "ORD395518" "ORD494479" "ORD019224" ...
 $ customer_id                 : chr  "ID6197211592" "ID0282825849" "ID0579391891" "ID4544561904" ...
 $ date                        : chr  "6/22/2019" "12/29/2019" "3/2/2019" "1/12/2019" ...
 $ nearest_warehouse           : chr  "Thompson" "Thompson" "Nickolson" "Nickolson" ...
 $ shopping_cart               : chr  "[('Lucent 330S', 1), ('Thunder line', 2), ('iStream', 2), ('pearTV', 1)]" "[('T
hunder line', 1), ('Universe Note', 2)]" "[('Thunder line', 1), ('pearTV', 2)]" "[('Universe Note', 1), ('Alcon 10',
2), ('Olivia x460', 1), ('iAssist Line', 1)]" ...
 $ order_price                 : int  12200 9080 10670 24800 9145 7810 13700 7960 25390 13320 ...
 $ delivery_charges            : num  79.9 62.7 65.9 57.6 75.5 ...
 $ customer_lat                : num  -37.8 -37.8 -37.8 -37.8 37.8 ...
 $ customer_long               : num  145 145 145 145 145 ...
 $ coupon_discount             : int  10 0 10 15 25 10 5 5 10 15 ...
 $ order_total                 : num  11060 9143 9669 21138 6934 ...
 $ season                      : chr  "Winter" "Summer" "Autumn" "Summer" ...
 $ is_expedited_delivery       : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
 $ distance_to_nearest_warehouse: num  1.28 1.162 1.095 0.857 0.587 ...
 $ latest_customer_review      : chr  "perfect phone and trusted seller. phone itself is amazing. i got this phone as
shown here on the description. i"| __truncated__ "it keeps dropping calls the wifi don't work this is a waste of money
this phone is worthless, it keeps dropping"| __truncated__ "five stars this is a great cheap phone." "charger did not f
it the charger didn't fit." ...
 $ is_happy_customer           : logi  TRUE FALSE TRUE FALSE TRUE FALSE ...
```

Hình 3.2. Cấu trúc dữ liệu của dirty_data

Dựa trên kết quả cho thấy, tập tin bao gồm 500 giá trị quan sát với 16 biến, trong đó các biến order_total, order_price, delivery_charges, coupon_discount, customer_lat distance_to_nearest_warehouse, customer_long, is_expedited_delivery là các biến định lượng còn lại là các biến định tính.

3.2 Xử lý dữ liệu

3.2.1. Khởi tạo dữ liệu

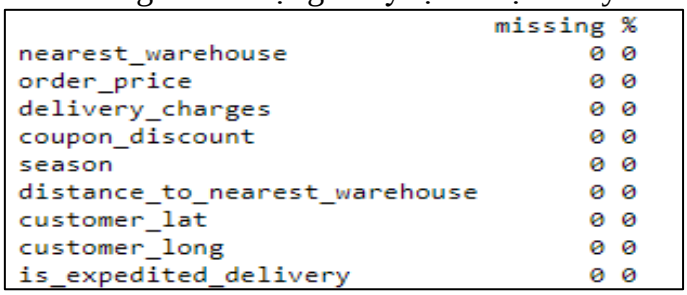
Trong quá trình phân tích dữ liệu, việc xác định và lựa chọn các biến phù hợp là rất quan trọng để nâng cao hiệu quả và tính chính xác của kết quả. Các biến order_id và id_customer không cung cấp được thông tin hữu ích cho phân tích. Với biến date trong trường trường hợp này là không cần thiết vì đã có biến seasons. Ở đề tài này tập trung vào việc phân tích chi phí vận chuyển. Do đó, chúng ta chỉ sử dụng biến order_price mà không sử dụng biến order_total vì order_total có liên quan tuyến tính mạnh đến order_price nên dễ xảy ra đa cộng tuyến khi đưa vào mô hình.

Ngoài ra, biến shopping_cart với dữ liệu từng biến khá phực tạp, gây khó khăn cho việc phân tích. Biến lastest_customer_review không phù hợp cho việc phân tích định lượng. Cuối cùng loại bỏ biến is_happy_customer vì biến này xảy ra sau khi có biến phí vận chuyển. Sau khi loại bỏ được các biến trên, ta tạo bộ dữ liệu mới gồm các biến: “nearest_warehouse, order_price, delivery_charges, coupon_discount,season, distance_to_nearest_warehouse, customer_lat, customer_long, s_expedited_delivery” .

	nearest_warehouse	order_price	delivery_charges	coupon_discount	season	distance_to_nearest_warehouse
1	Thompson	12200	79.89	10	Winter	1.2800
2	Thompson	9080	62.71	0	Summer	1.1621
3	Nickolson	10670	65.87	10	Autumn	1.0949
4	Nickolson	24800	57.61	15	Summer	0.8571
5	Nickolson	9145	75.54	25	Spring	0.5867
6	Bakers	7810	71.22	10	Spring	2.0752
7	Thompson	13700	74.84	5	Spring	0.6767
8	Thompson	7960	52.28	5	Autumn	1.3043
9	Bakers	25390	107.58	10	Spring	1.6595
10	Thompson	13320	62.26	15	winter	0.6093

Hình 3.3. Kết quả 10 dòng đầu tiên của new data

Tiếp theo đến việc kiểm tra dữ liệu khuyết, chúng ta sử dụng lệnh freq.na trong thư viện "questionr" để thống kê số lượng và tỷ lệ dữ liệu khuyết trong new_data



	missing	%
nearest_warehouse	0	0
order_price	0	0
delivery_charges	0	0
coupon_discount	0	0
season	0	0
distance_to_nearest_warehouse	0	0
customer_lat	0	0
customer_long	0	0
is_expedited_delivery	0	0

Hình 3.4. Số lượng và tỷ lệ dữ liệu khuyết trong new_data

Nhận thấy việc thống kê tổng dữ liệu khuyết ở các cột bằng 0, nên chúng ta chắc chắn rằng không tồn tại dữ liệu khuyết ở các biến, đảm bảo quá trình phân tích sau này.

3.2.2 Kiểm tra và đồng bộ dữ liệu

Chúng ta sử dụng biến unique để kiểm tra dữ liệu của các biến season và nearest_warehouse

```
> unique(new_data$season)
[1] "Winter" "Summer" "Autumn" "Spring" "winter" "spring" "autumn" "summer"
> unique(new_data$nearest_warehouse)
[1] "Thompson"  "Nickolson" "Bakers"    "nickolson" "thompson"
```

Hình 3.5. Dữ liệu các biến season và nearest_warehouse

Nhận xét: Kết quả cho thấy việc sai sót chính tả trong biến season và biến nearest_warehouse, nên chúng ta cần đồng bộ các dữ liệu lại thành "Winter", "Summer", "Spring", "Autum" cho biến season và "Thompson", "Nickolson", "Bakers" cho biến nearest_warehouse. Và sử dụng lệnh unique cho 2 biến để kiểm tra lại dữ liệu sau khi đồng bộ.

```
> unique(new_data$season)
[1] "Winter" "Summer" "Autumn" "Spring"
> unique(new_data$nearest_warehouse)
[1] "Thompson"  "Nickolson" "Bakers"
```

Hình 3.6. Kết quả đồng bộ của biến season và nearest_warehouse

Nhận xét: Kết quả cho thấy các biến trên đã đồng bộ về dữ liệu và có thể sử dụng cho khâu xử lý tiếp theo.

4. THỐNG KÊ MÔ TẢ

4.1. Thống kê mô tả cho các biến định lượng, định tính.

Để đảm bảo tính hiệu quả và độ chính xác trong quá trình phân tích đánh giá mô hình chúng ta cần đi xác định chính xác các biến định tính và định lượng trong dữ liệu.

Thực hiện tính toán thống kê mô tả cho các biến định lượng bao gồm: giá trị trung bình (mean), trung vị (median), độ lệch chuẩn (sd), giá trị nhỏ nhất (min), giá trị lớn nhất (max), mức phân vị thứ nhất (Q1), mức phân vị thứ ba (Q3). Kết quả thể hiện

ở Hình 4.1:

	order_price	delivery_charges	coupon_discount	order_total	distance_to_nearest_warehouse
mean	25522.22	76.65820	10.890000	39209.672	2.204224
median	12807.50	76.31000	10.000000	11293.960	1.030100
sd	86333.73	14.48146	8.649134	274194.016	8.812416
Q1.25%	7050.00	65.98250	5.000000	6454.735	0.751425
Q2.50%	12807.50	76.31000	10.000000	11293.960	1.030100
Q3.75%	20360.00	82.55500	15.000000	18119.188	1.408625
min	585.00	46.35000	0.000000	639.290	0.107800
max	947691.00	114.04000	25.000000	5688269.600	94.973400

Hình 4.1. Kết quả tính toán thống kê mô tả cho các biến định lượng

Thực hiện thống kê tần số cho các biến định tính cho biết số lượng các kho gần nhất, số lượng đơn hàng ở các mùa, tần số sử dụng dịch vụ giao hàng nhanh hay đánh giá trị sự hài lòng thể hiện ở Hình 4.2:

```
Bakers Nickolson  Thompson
   119       184       197
```

nearest_warehouse

```
Autumn Spring Summer Winter
   127    134    124    115
```

season

```
False  True
  251   249
```

is_expected_delivery

```
False  True
  141   359
```

is_happy_customer

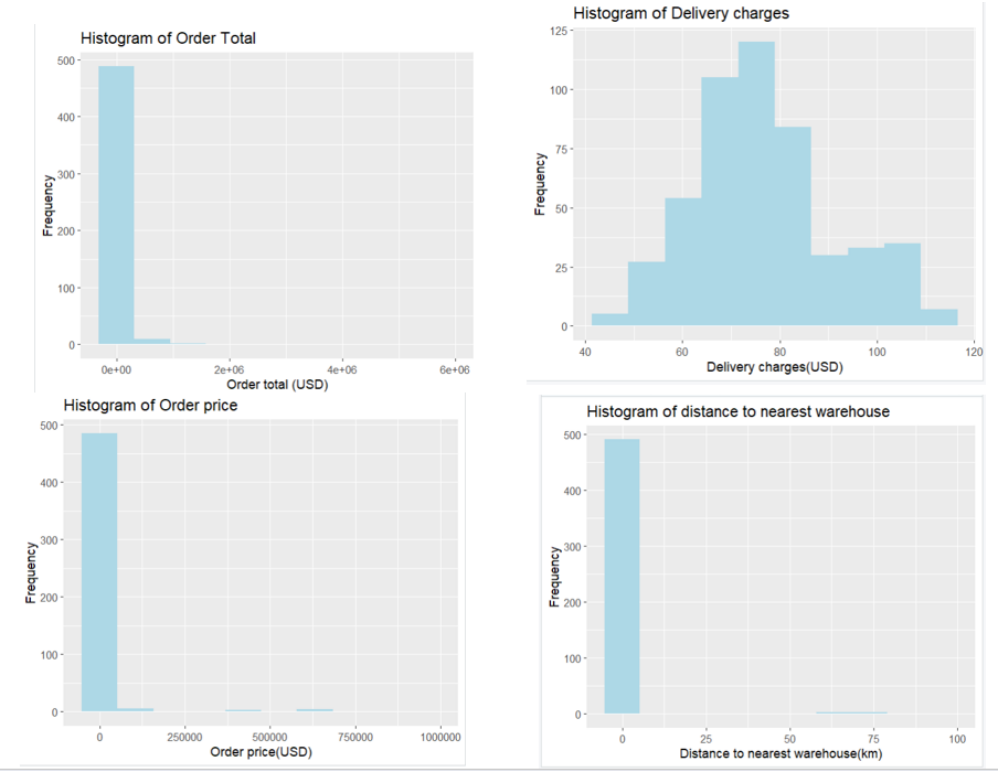
Hình 4.2. Kết quả khi lập bảng thống kê số lượng mỗi phân loại

4.2. Mô tả phân phối dữ liệu bằng các đồ thị

4.2.1. Biểu diễn sự phân phối của các biến định lượng thông qua đồ thị Histogram

Histogram, hay biểu đồ tần suất, là một loại biểu đồ cột dùng để thể hiện sự phân bố dữ liệu theo các khoảng giá trị. Mỗi cột trong biểu đồ phản ánh tần suất xuất hiện của dữ liệu trong khoảng giá trị tương ứng. Histogram giúp chúng ta phân tích các đặc điểm của dữ liệu, như mức độ tập trung ở một khoảng giá trị nhất định, xu hướng lệch, sự phân tán, cũng như phát hiện các giá trị bất thường (outliers) trong dữ liệu.

Xét đồ thị Histogram của các biến định lượng liên tục:



Hình 4.3. Biểu đồ Histogram của các biến định lượng liên tục

Nhận xét: Từ Hình 4.3, chúng ta nhận thấy đồ thị histogram của biến order_total, order_price, distance_nearest_warehouse có phân phối lệch mạnh về phía trái, có duy nhất biến delivery_charges gần giống với phân phối dạng chuẩn hình chuông úp ngược. Nhận xét sâu hơn về các đồ thị, ở đồ thị order_total, và order_price phần lớn đều có giá trị thấp gần về trục tung trong khi đó có rất ít đơn hàng giá trị lớn, ở order_total có giá trị lớn như 6e+06 hay ở order_price lại có giá trị này nằm rải rác trong khoảng 380000-1000000. Sự xuất hiện của một số giá trị lớn bất thưởng nà làm kéo dài trục x, dẫn đến hiện tượng giá trị tập trung về phía bên trái, đây cũng có thể là điểm ngoại lai khiến trở nên mất đối xứng. Chúng ta cũng có thể dễ dàng nhận thấy giá trị bất thường cũng xuất hiện ở đồ thị distance_neareast_warehouse ở khoảng 60 đến 75km, còn lại phần đa đều tập trung ở giá trị thấp.

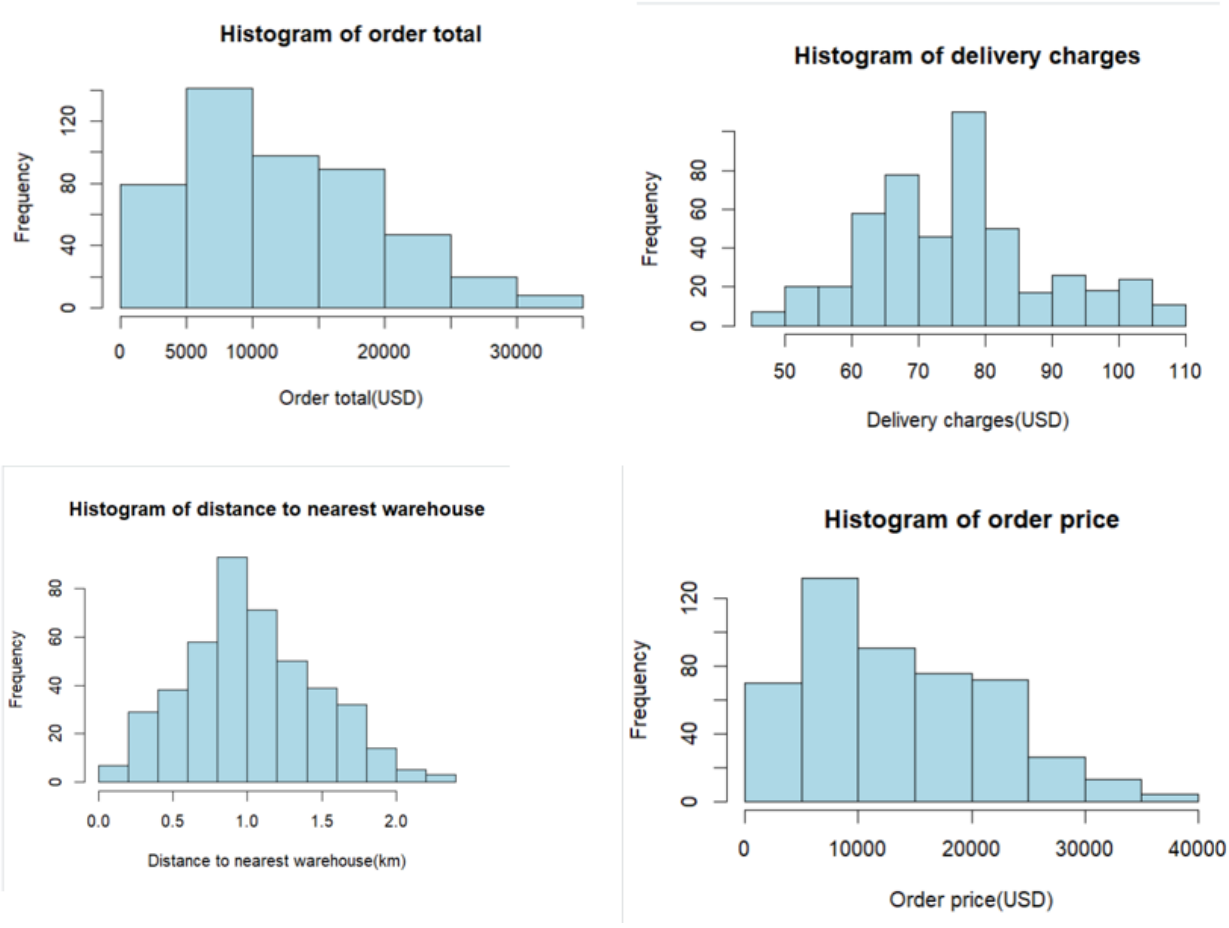
Do đó, chúng ta cần xác định tỷ lệ giá trị ngoại lai trong các biến và đánh giá xem việc loại bỏ chúng có ảnh hưởng đến độ chính xác trong quá trình đánh giá và phân tích mô hình hay không. Thực hiện bằng code R tạo hàm remove_outliers ở dòng 81, và đếm giá trị ngoại lai, việc xác định ngoại lai bằng phương pháp IQR, với IQR (IQR=Q3-Q1) là sự khác biệt giữa tứ phân vị thứ nhất Q1 và tứ phân vị thứ ba Q3 dùng để thiết lập các giá trị biên. Giá trị được gọi là ngoại lai (Outlier) nếu nhỏ hơn Q1-1.5×IQR và lớn hơn Q3+1.5×IQR.

```
                                  missing %
order_total                            18 4
order_price                            16 3
distance_to_nearest_warehouse          12 2
nearest_warehouse                       0 0
delivery_charges                        0 0
coupon_discount                         0 0
season                                  0 0
is_expedited_delivery                   0 0
is_happy_customer                       0 0
> |
```

Hình 4.4. Thống kê các giá trị ngoại lai

Nhận xét: Dựa trên bảng thống kê hình 4.4, các giá trị ngoại lai chiếm tỷ lệ nhỏ trong mỗi biến. Cụ thể, với biến order_total, tổng cỡ mẫu là 500, trong đó có 18 giá trị ngoại lai, chiếm 3,6% (xấp xỉ 4%). Cũng như biến order_price và distance_to_nearest_warehouse có ngoại lai chiếm lần lượt 3% và 2% trong tổng cỡ mẫu. Do đó việc loại bỏ giá trị ngoại lai là phù hợp vẫn đảm bảo được tính chính xác trong quá trình phân tích và đánh giá. Tạo một tệp tin new_data2 lưu dữ liệu sau khi xóa bỏ ngoại lai được thực hiện trong code R dòng 122.

Tiến hành xóa bỏ ngoại lai:



Hình 4.5. Biểu đồ Histogram của các biến định lượng sau khi xóa bỏ ngoại lai

Nhận xét: Từ Hình 4.5, nhìn chung, hầu hết các đồ thị đều thể hiện sự đối xứng và có hình dáng gần giống với "hình chuông úp ngược" đặc trưng của phân phối chuẩn. Điều này cho thấy các biến sau khi xử lý đã có phân phối xấp xỉ chuẩn, phù hợp để áp dụng các phương pháp phân tích thống kê dựa trên giả định phân phối chuẩn.

4.2.2. Biểu diễn sự phân phối của các biến định lượng theo order_total thông qua biểu đồ hộp Boxplot

Biểu đồ hộp (Box plot) hay còn gọi là biểu đồ hộp và râu (Box and whisker plot) là biểu đồ diễn tả 5 vị trí phân bố của dữ liệu, đó là: giá trị nhỏ nhất (min), tứ phân vị thứ nhất (Q1), trung vị (median), tứ phân vị thứ 3 (Q3) và giá trị lớn nhất (max). Nó thường được sử dụng so sánh phân phối của các nhóm dữ liệu thông qua phạm vi của hộp và râu của biểu đồ.

Phân tích mối quan hệ giữa biến order_total và is_happy_customer

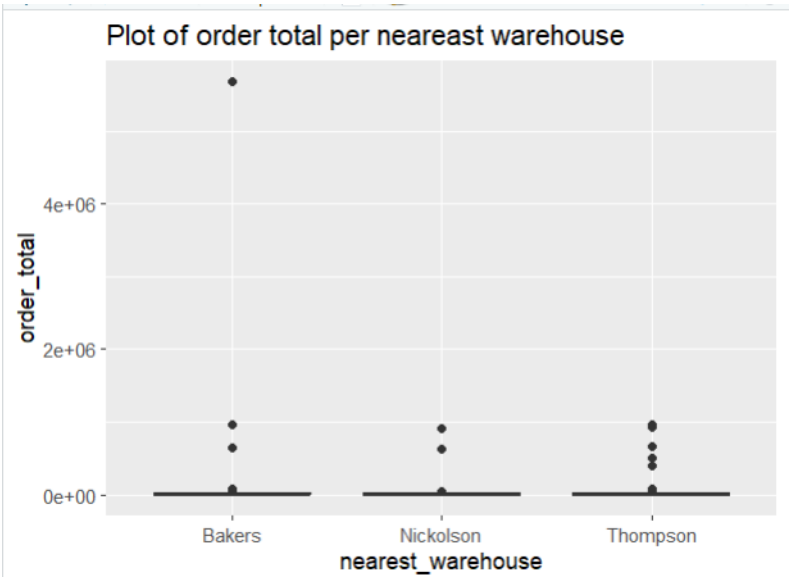
Hình 4.6. Đồ thị Boxplot thể hiện phân phối của order_total theo is_happy_customer

Nhận xét: Từ Hình 4.6 trên, chúng ta có thể thấy rằng mức trung vị (median) của order_total nằm gần cạnh dưới của hộp, chỉ ra rằng phần lớn các giá trị tập trung ở mức thấp. Độ lệch chuẩn mẫu giữa hai nhóm khách hàng hài lòng (True) và khách hàng không hài lòng (False) tương đối nhỏ, cho thấy sự đồng nhất về mức giá trị trong mỗi nhóm. Tuy nhiên, nhìn nhận rõ cả hai nhóm đều có các giá trị ngoại lai (outliers) rõ rệt nằm ngoài râu của hộp, cách biệt đáng kể so với các giá trị còn lại. Các điểm ngoại lai này có thể ảnh hưởng đến kết quả phân tích, vì vậy cần phải xóa bỏ ngoại lai bằng phương pháp IQR được cập nhật trong new_data2.

Hình 4.7. Đồ thị Boxplot thể hiện phân phối của order_total theo is_happy_customer sau khi xóa bỏ ngoại lai.

Nhận xét: Từ Hình 4.7, chúng ra rút ra nhận xét: sau khi loại bỏ các giá trị ngoại lai, cả hai nhóm dữ liệu True và False của biến is_happy_customer trở nên tập trung hơn và đồng đều hơn. Việc so sánh giữa hai nhóm trở nên dễ dàng hơn, vì sự phân bố của các giá trị order_total trong mỗi nhóm có sự đồng nhất rõ rệt hơn. Mức trung bình mẫu của hai nhóm khách hàng hài lòng (True) và khách hàng không hài lòng (False) gần như tương đương, điều này cho thấy rằng sự hài lòng của khách hàng không có ảnh hưởng đáng kể đến giá trị order_total sau khi loại bỏ các ngoại lai. Như vậy, việc xóa bỏ ngoại lai đã giúp làm rõ hơn mối quan hệ giữa các biến và làm tăng độ chính xác của phân tích.

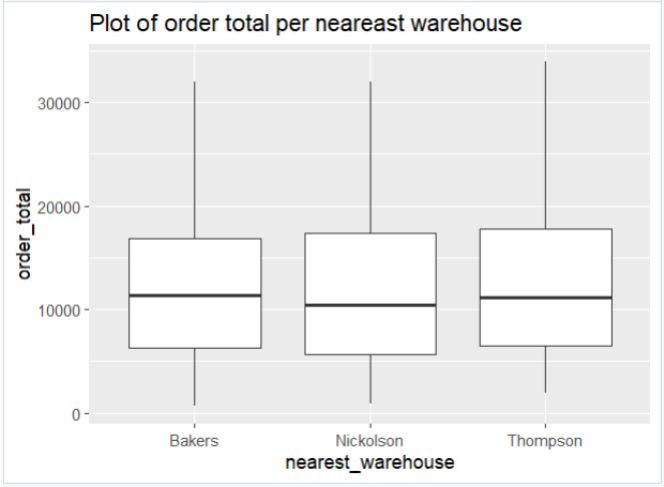
Thực hiện kiểm định các giá trị sản phẩm order_total tương ứng với mỗi nearest_warehouse



Hình 4.8. Đồ thị Boxplot thể hiện phân phối của order_total theo nearest_warehouse.

Nhận xét: Từ biểu đồ hộp Hình 4.8, chúng ta thấy rằng được chia thành ba nhóm:

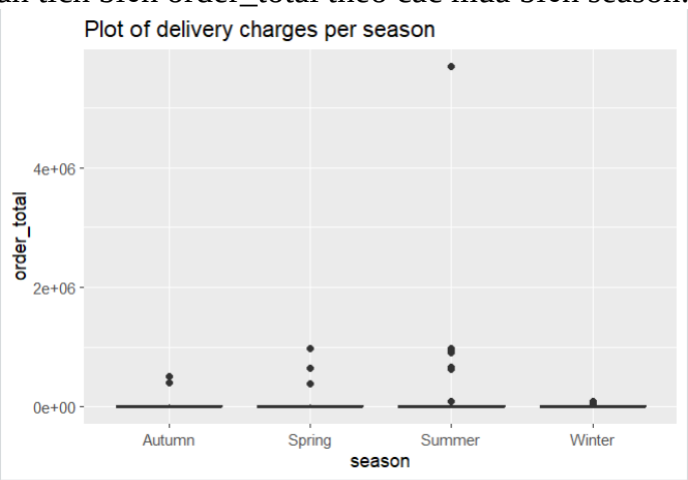
Bakers, Nickosol, Thompson của biến nearest_warehouse. Cả ba nhóm đều có mức trung vị order_total nằm rất gần nhau là gần đáy của phối, cho thấy phần lớn đơn hàng có giá trị thấp hơn rất nhiều so với các giá trị ngoại lai ở ngoài râu hộp, gây ảnh hưởng đáng kể đến phân tích khiến việc đánh giá mức độ khác biệt giữa các kho hàng thiếu chính xác. Việc loại bỏ ngoại lai sẽ giúp dữ liệu phản ánh đúng hơn thực trạng.



Hình 4.9. Biểu đồ Boxplot thể hiện phân phối của order_total theo nearest_warehouse sau khi xóa bỏ ngoại lai

Nhận xét: Từ Hình 4.9, sau khi loại bỏ các giá trị ngoại lai, cả ba nhóm dữ liệu đều có phân phối giá trị trung bình order_total tương tự nhau cho thấy các kho hoạt động khá đồng đều về giá trị đơn hàng. Trong đó, Thompson có giá trị order_total cao hơn so với hai cửa hàng còn lại, nhưng sự chênh lệch này không quá lớn.

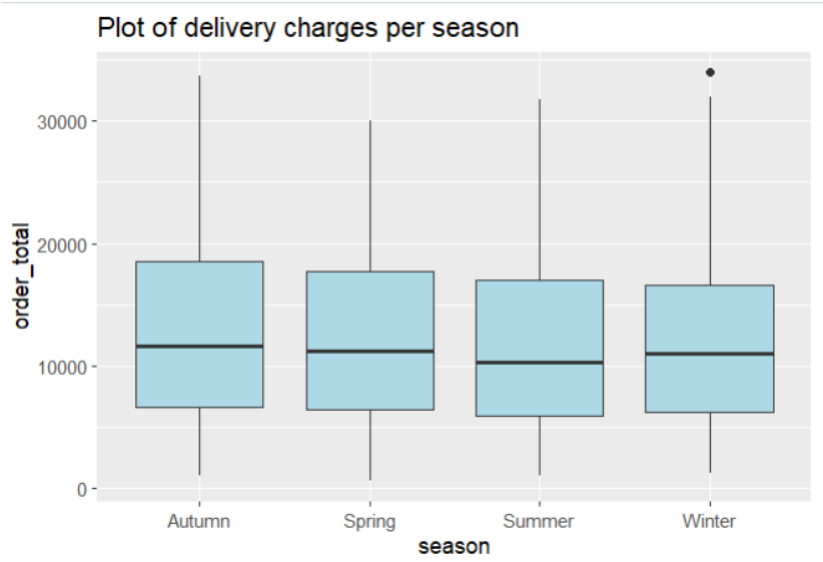
Thực hiện phân tích biến order_total theo các mùa biến season:



Hình 4.10. Biểu đồ Boxplot thể hiện phân phối của order_total theo season

Nhận xét: Biểu đồ Hình 4.10, thể hiện sự phân tán dữ liệu rất mạnh, đặc biệt ở mùa hè (Summer) với một số giá trị vượt mức 4 triệu. Các giá trị này nằm ngoài râu hộp, tạo thành các điểm ngoại lai rõ rệt, trong khi phần lớn dữ liệu bị nén sát trục dưới. Điều này khiến giá trị trung bình bị kéo xuống gần đáy, làm nổi bật sự chênh lệch giữa các giá trị ngoại lai và phần lớn dữ liệu thực. Những giá trị bất thường này gây ảnh hưởng lớn đến phân tích, và xu hướng chính của tập dữ liệu.

Tiến hành xóa bỏ ngoại lai



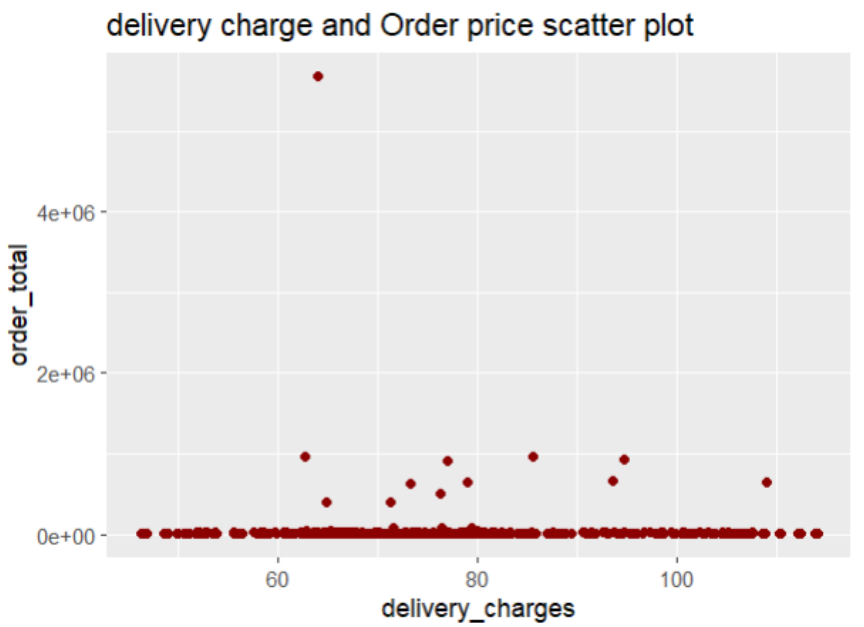
Hình 4.11. Biểu đồ Boxplot thể hiện phân phối của order_total theo season sau khi xóa bỏ ngoại lai

Nhận xét: Từ hình 4.11 trên, chúng ta thấy rằng dữ liệu sau khi xóa ngoại lai được biểu diễn dưới dạng boxplot có phân phối cân đối hơn, có mức trung vị của các mua mùa dao động từ 10.000 đến 15.000 nằm gần nhau cho thấy không có sự khác biệt quá lớn giữa các mùa.

4.2.3. Mô tả mối quan hệ giữa các biến bằng đồ thị scatter plot

Đồ thị scatter plot (biểu đồ phân tán) là một công cụ trực quan mạnh mẽ để biểu diễn mối quan hệ giữa hai biến số (tuyến tính, phi tuyết, không có mối quan hệ nào). Mỗi điểm trên đồ thị đại diện cho một cặp giá trị (x, y), trong đó trục hoành (x) thể hiện giá trị của một biến và trục tung (y) thể hiện giá trị của biến còn lại.

Đồ thị thể hiện mối quan hệ giữa order_total và delivery_charges

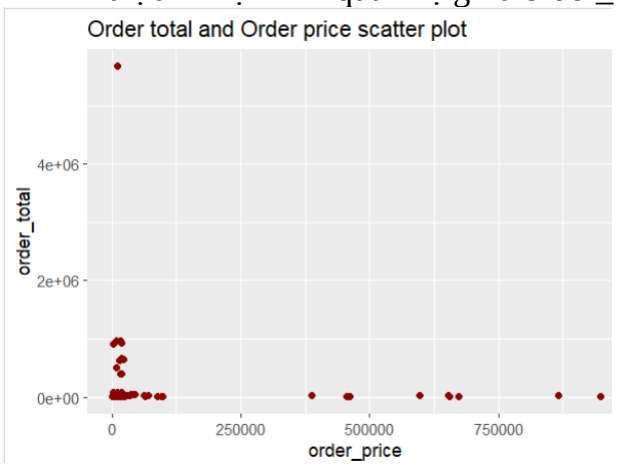


Hình 4.12. Biểu đồ phân tán giữa deliver_charges và order_total.

Nhận xét: Từ đồ thị Hình 4.12 trên, chúng ta không thấy được xu hướng rõ ràng hay

quan hệ tuyến tính yếu giữa "delivery_charges" và "order_total". Điều này có thể thể hiện rằng phí giao hàng không thực sự ảnh hưởng nhiều đến giá trị tổng của đơn hàng. Điểm ngoại lai không ảnh hưởng đến đồ thị, đồ thị vẫn thể hiện được sự phân bố tập trung nên không cần xử lí ngoại lai.

Đồ thị thể hiện mối quan hệ giữa order_total và order_price



Hình 4.13: Mối quan hệ giữa order_price và order_total chưa qua xử lí ngoại lai.

Hình 4.14: Mối quan hệ giữa order_price và order_total sau khi xử lí ngoại lai.

Nhận xét: Do đồ thị Hình 4.13 chưa xử lí ngoại lai có nhiều điểm bất thường so với điểm khác, có giá trị rất cao so với giá trị khác, ngoài ra điểm ngoại lai có thể là do sai số khiến cho đồ thị giãn ra rất nhiều. Vì vậy cần xóa bỏ ngoại lai để có thể thu được đồ thị tập trung và phân bố đều hơn như thể hiện ở hình 4.14 để phân tích và được thể hiện qua hình bên phải, từ đó có thể nhận xét rằng giữa order_total có mối quan hệ tuyến tính mạnh với order_price.

5. THỐNG KÊ SUY DIỄN

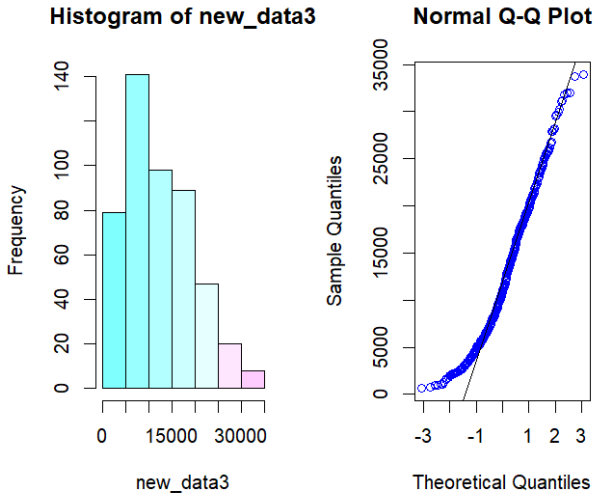
5.1. Mô hình ước lượng trung bình một mẫu

Bài toán: Với mức ý nghĩa 5%, thực hiện ước lượng trung bình tổng chi phí đặt

hàng order_total.

Bước 1: Kiểm tra xem order_total có tuân theo phân phối chuẩn hay không

Vẽ đồ thị: Chúng ta thực hiện vẽ đồ thị Histogram và đồ thị Q_Q Plot của order_total để xem nó có tuân theo phân phối chuẩn.



Hình 5.1. Đồ thị Histogram và đồ thị Q_Q Plot của order_total

Nhận xét:

- Qua đồ thị Histogram ở Hình 5.1: Chúng ta thấy dữ liệu không có dạng hình chuông hay đối xứng nên ta nghi ngờ rằng order_total không tuân theo quy luật phân phối chuẩn.
- Qua đồ thị Q_Q Plot ở Hình 5.1: Chúng ta thấy có các điểm quan trắc không nằm trên đường thẳng kỳ vọng phân phối chuẩn nên chúng nghi ngờ rằng order_total không tuân theo quy luật phân phối chuẩn. Tuy nhiên, việc kiểm tra một biến có tuân theo phân phối chuẩn hay không bằng cách vẽ đồ thị chỉ mang tính chất cảm quan, tham khảo và có thể sẽ không được chính xác. Để có kết luận chính xác hơn chúng ta kiểm định lại bằng hàm Shapiro – Wilk.

Kiểm định bằng hàm Shapiro – Wilk: để có thể kết luận chính xác về việc order_total có tuân theo phân phối hay không chúng ta định kiểm định bằng hàm Shapiro – Wilk bằng câu lệnh chính là shapiro.test(x) trong code Rstudio .

- Đặt giả thuyết kiểm định:
  + Giả thuyết H0: Tổng chi phí đặt hàng tuân theo phân phối chuẩn.
  + Đối thuyết H1: Tổng chi phí đặt hàng không tuân theo phân phối chuẩn.

```
        Shapiro-Wilk normality test

data:  new_data3
W = 0.95647, p-value = 9.958e-11
```

Hình 5.2. Kết quả kiểm định Shapiro – Wilk cho order_total

Nhận xét: Sau khi thực hiện kiểm định bằng hàm Shapiro – Wilk, chúng ta nhận được kết quả trong Hình 5.2 với giá trị p-value = 9.958×10^{-11} nhỏ hơn mức ý nghĩa 5% nên ta đủ điều kiện bác bỏ giả thuyết H0, chấp nhận đối thuyết H1.

Vậy tổng chi phí đặt hàng order_total không tuân theo quy luật phân phối chuẩn.

Bước 2: Xác định loại công thức phù hợp cho bài toán

— Chưa biết phương sai của tổng thể σ^2.

— Có kích thước cỡ mẫu: n ≥ 30.

— Không tuân theo quy luật phân phối chuẩn (có phân phối tùy ý).

Bước 3: Tính toán các giá trị cần thiết để tìm khoảng tin cậy: Để tìm khoảng tin cậy của order_total ta cần tìm các giá trị như tổng thể, giá trị trung bình, độ lệch chuẩn, điểm phân vị mức của phân phối chuẩn hóa với mức ý nghĩa 5%, từ đó chúng ta tìm được độ chính xác và tìm được khoảng tin cậy của biến order_total.

	Tong_the	Trung_binh	Do_lech_chuan	Z	Do_chinh_xac	Can_duoi	Can_tren
1	482	12206.42	7265.276	1.959964	648.5997	11557.82	12855.02

Hình 5.3. Kết quả ước lượng trung bình order_total

Nhận xét: Từ các giá trị tính được trong Hình 5.3,chúng ta xác đinh được khoảng tin cậy 95% cho trung bình tổng thể là:

Khoảng tin cậy = (12206.42 – 648.60, 12206.42 + 648.60) = (11557.82, 12855.02).

5.2. Mô hình kiểm định trung bình hai mẫu

Bài toán: Với mức ý nghĩa 5%, sự hài lòng của khách hàng với đơn hàng (is_happy_customer) có bị ảnh hưởng đến chi phí đặt hàng (order_total) hay không?

- Gọi μ_1: Chi phí đặt hàng trung bình của khách hàng đánh giá hài lòng.
- Gọi μ_2: Chi phí đặt hàng trung bình của khách hàng đánh giá không hài lòng.

Bước 1: Đặt giả thuyết:

— Giả thuyết H0: $\mu_1 = \mu_2$.

— Đối thuyết H1: $\mu_1 < \mu_2$.

Bước 2: Tạo tệp tin gồm 2 biến is_happy_customer và order_total để thực hiện kiểm định 2 mẫu. Sử dụng hàm phân tích ngoại lai và xóa các hàng chứa giá trị ngoại lai và xuất bảng giá trị xem xét các số lượng còn lại của mỗi phần tử sau khi loại bỏ ngoại lai. Sau đó, tách biến định tính is_happy_customer và loại bỏ ngoại lai từ dữ liệu gốc.

```
False  True
  136   346
```

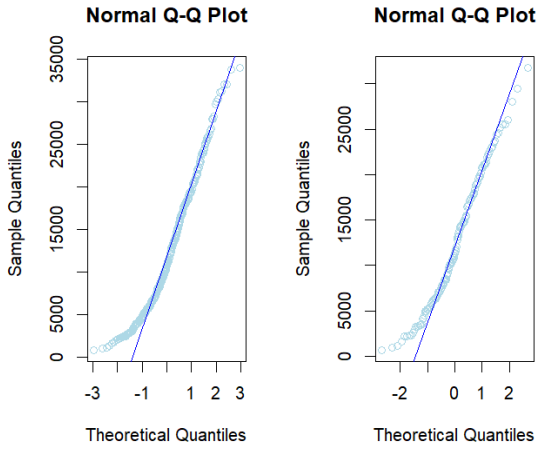
Hình 5.4. Kết quả khi loại bỏ ngoại lai của 2 nhóm trong biến is_happy_customer

Nhận xét: Từ kết quả trên Hình 5.4, chúng ta thấy rằng sau khi loại bỏ ngoại lai có 136

khách hàng cảm thấy hài lòng với chi phí đặt hàng của cửa hàng và 346 khách hàng cảm thấy không hài lòng với chi phí đặt hàng của cửa hàng.

Bước 3: Kiểm tra xem order_total của TRUE_data và FALSE_data có tuân theo phân phối chuẩn hay không.

Vẽ đồ thị Q_Q Plot của order_total trong hai tập tin TRUE_data và FALSE_data để xem chúng có tuân theo phân phối chuẩn hay không.



Hình 5.5. Đồ thị Q_Q Plot của order-price trong TRUE-data (bên trái) và FALSE-data (bên phải)

Nhận xét: Từ hai đồ thị Q_Q Plot trong Hình 5.5 cho chúng ta thấy phần lớn các điểm nằm trên đường kỳ vọng phân phối chuẩn, nhưng có một số điểm lệch ra khỏi đường này, đặc biệt ở các điểm đầu và cuối ở cả hai đồ thị. Điều này xác nhận rằng dữ liệu không hoàn toàn tuân theo phân phối chuẩn. Tuy nhiên, việc kiểm tra một biến có tuân theo phân phối chuẩn hay không bằng cách vẽ đồ thị chỉ mang tính chất cảm quan, tham khảo và có thể sẽ không được chính xác. Để có kết luận chính xác hơn chúng ta kiểm định lại bằng hàm Shapiro – Wilk.

Kiểm định bằng hàm Shapiro – Wilk: để có thể kết luận chính xác về việc order_total trong hai tập tin TRUE_data và FALSE_data có tuân theo phân phối hay không chúng ta định kiểm định bằng hàm Shapiro – Wilk bằng câu lệnh chính là shapiro.test(x) trong code Rstudio.

- Đặt giả thuyết kiểm định:

— Nhóm khách hàng hài lòng:

+ H0: Chi phí đặt hàng của nhóm khách hàng hài lòng tuân theo phân phối chuẩn.
+ H1: Chi phí đặt hàng của nhóm khách hàng hài lòng không tuân theo phân phối chuẩn.

— Nhóm khách hàng không hài lòng:

+ H0: Chi phí đặt hàng của nhóm khách hàng không hài lòng tuân theo phân phối chuẩn.
+ H1: Chi phí đặt hàng của nhóm khách hàng không hài lòng không tuân theo phân phối chuẩn.

TRUE_data FALSE_data

Hình 5.6. Kết quả kiểm định Shapiro – Wilk cho order_total trong 2 nhóm.

Nhận xét: Sau khi thực hiện kiểm định bằng hàm Shapiro – Wilk, chúng ta nhận được kết quả trong Hình 5.6 với các giá trị p-value của mỗi nhóm đều nhỏ hơn mức ý nghĩa 5% nên ta đủ điều kiện bác bỏ giả thuyết H0, chấp nhận đối thuyết H1 ở cả hai nhóm. Vậy order_total ở cả hai tập đều không tuân theo phân phối chuẩn.

Bước 4: Xác định công thức

— Kích thước hai mẫu đều lớn hơn 30

— Không tuân theo quy luật phân phối chuẩn (có phân phối tùy ý).

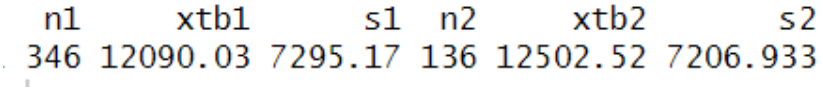
Bước 5: Thức hiện kiểm định

- Đặt giả thuyết kiểm định:

— Giả thuyết H0: Sự hài lòng của khách hàng không bị ảnh hưởng bởi chi phí đặt hàng.

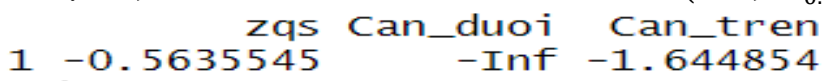
— Đối thuyết H1: Sự hài lòng của khách hàng không bị ảnh hưởng bởi chi phí đặt hàng.

Tính các giá trị cần thiết để kiểm định thống kê của mỗi biến: tổng thể mẫu, trung bình mẫu, độ lệch chuẩn mẫu.



Hình 5.7. Giá trị kiểm định

- Khi mẫu lớn hơn 30, phân phối của thống kê kiểm định phân phôi chuẩn, do đó ta dùng kiểm định z, do xtb1 < xtb2 nên miền bác bỏ RR = $(-\infty,-z_{0.05})$.



Hình 5.8. Kết quả kiểm định z

Nhận xét: Từ kết quả kiểm định z trên Hình 5.7, chúng ta thấy giá trị kiểm định z_{qs}= -0.56 không thuộc miền bác bỏ $RR = (-\infty,-1.65)$ chưa thể bác bỏ giả thuyết H0. Vậy sự hài lòng của khách hàng không bị ảnh hưởng bởi chi phí đặt hàng.

- Giả sử hai mẫu có phân phối chuẩn ta sử dụng kiểm định *t*

— Thực hiện kiểm định phương sai hai mẫu: bằng câu lệnh chính là var.test(x,...).

+ Giả thuyết H0: Phương sai hai mẫu giống nhau ($\sigma_1^2 = \sigma_2^2$)
+ Đối thuyết H1: Phương sai của TRUE_data nhỏ hơn FALSE_data ($\sigma_1^2 < \sigma_2^2$)

```
        F test to compare two variances

data:  TRUE_data$order_total and FALSE_data$order_total
F = 1.0246, num df = 345, denom df = 135, p-value = 0.5589
alternative hypothesis: true ratio of variances is less than 1
95 percent confidence interval:
 0.00000 1.28924
sample estimates:
ratio of variances
          1.024637
```

Hình 5.9. Kết quả kiểm định phương sai bằng var.test

Nhận xét: Từ kết quả trên Hình 5.9, chúng ta rút ra ta có p-value = 0.5589 lớn hơn mức ý nghĩa 5% nên chưa đủ điều kiện để bác bỏ giả thuyết H0. Vậy phương sai của 2 mẫu là giống nhau.

- Thực hiện kiểm định t

```
     sp_2        tqs Can_duoi  Can_tren
52859604 -0.5605685     -Inf -1.648034
```

Hình 5.10. Kết quả kiểm định t

Nhận xét: Từ kết quả kiểm định trên Hình 5.10, chúng ta thấy giá trị kiểm định $t_{qs} = -0.56$ không thuộc miền bác bỏ $RR = (-\infty, -1.65) \Rightarrow$ chưa đủ điều kiện bác bỏ giả thuyết H0. Vậy sự hài lòng của khách hàng không bị ảnh hưởng bởi chi phí đặt hàng.
⇒***Kết luận:*** Sự hài lòng của khách hàng không bị ảnh hưởng bởi chi phí đặt hàng.

5.3. Mô hình ANOVA một nhân tố

Bài toán: Phân tích sự ảnh hưởng của các kho hàng nearest_warehouse đến giá trị đơn đặt hàng cuối cùng order_total.

Tạo một tệp tin gồm 2 biến nearest_warehouse và order_total để thực hiện phân tích phương sai (ANOVA) một nhân tố. Sử dụng hàm phân tích ngoại lai và xóa các hàng chứa giá trị ngoại lai và xuất bảng giá trị xem xét các số lượng còn lại của mỗi phần tử sau khi loại bỏ ngoại lai).

```
Bakers Nickolson  Thompson
   113        181        188
```

Hình 5.11. Kết quả khi loại bỏ ngoại lai của 3 nhóm trong biến nearest_warehouse

Nhận xét: Dựa trên kết quả ở Hình 5.11, chúng ta thấy rằng còn 113 khách hàng ở kho Bakers ảnh hưởng tới chi phí đặt hàng, 181 khách hàng ở kho Nickolson ảnh hưởng tới chi phí đặt hàng và 188 khách hàng ở kho Thompson ảnh hưởng tới chi phí đặt hàng.
Để thực hiện ANOVA ta cần tách ba nhóm (Bakers, Thompson và Nickolson) của nearest_warehouse trong tệp tin vừa tạo (dữ liệu thực hiện ANOVA đã được tiền xử lý và xử lý ngoại lai) và sau đó kiểm tra các điều kiện cần để thực hiện.

5.3.1. Điều kiện thực hiện ANOVA

***Điều kiện 1*:** Kiểm tra phân phối chuẩn

Ta đặt giả thuyết kiểm định thống kê

- Đối với kho Bakers

— Giả thuyết H0: Tổng chi phí đặt hàng trung bình của kho Bakers tuân theo phân phối chuẩn.
— Giả thuyết H1: Tổng chi phí đặt hàng trung bình của kho Bakers không tuân theo phân phối chuẩn.

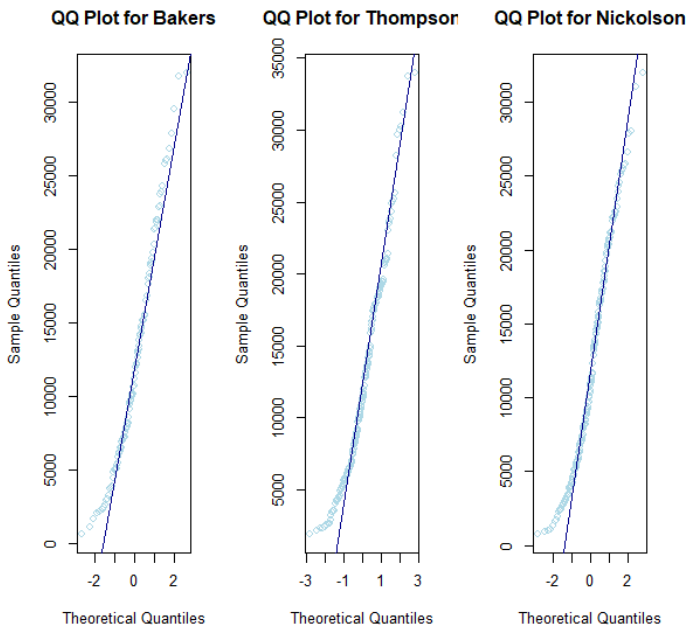
- Đối với kho Thompson

— Giả thuyết H0: Tổng chi phí đặt hàng trung bình của kho Thompson tuân theo phân phối chuẩn.
— Giả thuyết H1: Tổng chi phí đặt hàng trung bình của kho Thompson không tuân theo phân phối chuẩn.

- Đối với kho Nickolson

— Giả thuyết H0: Tổng chi phí đặt hàng trung bình của kho Nickolson tuân theo phân phối chuẩn.
— Giả thuyết H1: Tổng chi phí đặt hàng trung bình của kho Nickolson không tuân theo phân phối chuẩn.

Vẽ đồ thị của ba kho: Vẽ đồ thị Q_Q Plot của order_total ở từng kho hàng để xem chúng có tuân theo phân phối chuẩn hay không.



Hình 5.12. Đồ thị Q_Q plot của ba kho

Nhận xét: Dựa vào ba đồ thị Q_Q Plot trên Hình 5.12, chúng ta thấy các điểm quan trắc nằm lệch ra khỏi đương thẳng kỳ vọng phân phối chuẩn, nên chúng ta có thể dự đoán rằng tổng chi phí đặt hàng ở các kho hàng không tuân theo phân phối chuẩn. Tuy nhiên, việc kiểm tra một biến có tuân theo phân phối chuẩn hay không bằng cách vẽ đồ thị chỉ mang tính chất cảm quan, tham khảo và có thể sẽ không được chính xác. Để có kết luận chính xác hơn chúng ta kiểm định lại bằng hàm Shapiro – Wilk.

Kiểm định bằng hàm Shapiro – Wilk: để có thể kết luận chính xác về việc tổng chi phí đặt hàng ở các kho hàng có tuân theo phân phối hay không chúng ta định kiểm định bằng hàm Shapiro – Wilk bằng câu lệnh chính là shapiro.test(x) trong code Rstudio .

```
>  shapiro.test(Bakers$order_total)

        Shapiro-Wilk normality test

data:  Bakers$order_total
W = 0.95493, p-value = 0.0007802

>   shapiro.test(Nickolson$order_total)

        Shapiro-Wilk normality test

data:  Nickolson$order_total
W = 0.95366, p-value = 1.182e-05

>   shapiro.test(Thompson$order_total)

        Shapiro-Wilk normality test

data:  Thompson$order_total
W = 0.94607, p-value = 1.587e-06
```

Hình 5.13. Kết quả kiểm định Shapiro – Wilk

Nhận xét: Sau khi thực hiện kiểm định bằng hàm Shapiro – Wilk, chúng ta nhận được kết quả trong Hình 5.13, với tất cả giá trị p-value của 3 kho đều nhỏ hơn mức ý nghĩa là 5% nên ta đủ điều kiện bác bỏ giả thuyết H0, chấp nhận đối thuyết H1. Vậy tổng chi phí đặt hàng trung bình của các kho hàng không tuân theo phân phối chuẩn.

Điều kiện 2: Kiểm tra sự đồng nhất phương sai: Để kiểm tra sự đồng nhất về phương sai chúng ta dùng hàm Levene's Test kiểm tra sự đồng nhất phương sai của ba kho trong nearest_warehouse.

Đặt giả thuyết kiểm định thống kê:

— Giả thuyết H0: $\mu_1 = \mu_2 = \mu_3$.

— Đối thuyết H1: $\exists \mu_i \neq \mu_j$, với $1 \leq i, j \leq 3$ và $i \neq j$.

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2   0.771 0.4631
      479
```

Hình 5.14. Kết quả kiểm định Levene's Test

Nhận xét: Từ kết quả nhận được ở Hình 5.14, chúng ta thấy được p-value=0.4631 lớn hơn mức ý nghĩa 5% nên chúng ta chưa đủ điều kiện bác bỏ giả thuyết H0.
Vậy không có sự khác biệt về phương sai tổng chi phí đặt hàng của các kho.

Điều kiện 3: Các mẫu được lấy là độc lập

Nhận xét: Các giá trị trong cột order_total được lấy một cách độc lập. Mỗi đơn hàng không phụ thuộc vào các đơn hàng khác, và chúng được kết nối với các nhà kho gần nhất dựa trên vị trí hoặc một tiêu chí nhất định.

5.3.2. Thực hiện ANOVA

Thực hiện ANOVA một nhân tố để so sánh tổng chi phí đặt hàng trung bình giữa ba kho hàng.

- Đặt giả thuyết kiểm định thống kê:

— Giả thuyết H0: $\mu_1 = \mu_2 = \mu_3$: Tổng chi phí đặt hàng trung bình giữa ba kho là tương đương nhau.

— Đối thuyết H1: $\exists \mu_i \neq \mu_j$, với $1 \leq i, j \leq 3$ và $i \neq j$: Tồn tại ít nhất hai kho hàng có sự khác biệt về tổng chi phí đặt hàng trung bình.

- Bảng thống kê các giá trị ở ba kho: Chúng ta đi tìm các số lượng khách mua hàng ở từng kho và tổng trong binhg chi phí đặt hàng ở từng kho

	Warehouse	n	Trung_binh
1	Bakers	113	12303.72
2	Nickolson	181	11970.55
3	Thompson	188	12375.01

Hình 5.15. Kết quả tính toán giá trị thống kê cơ bản

- Tiến hành ANOVA: ta tiến hành ANOVA của hai biến order_total và nearest_warehouse bằng câu lệnh chính là aov(...) trong Rstudio (xem dòng code 266 và 267 trong phần Link code ở chương số 7).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
nearest_warehouse	2	1.648e+07	8241726	0.156	0.856
Residuals	479	2.537e+10	52970224		

Hình 5.16. Kết quả của mô hình ANOVA

Nhận xét: Từ kết quả trên Hình 5.16, ta có bảng sau:

Source of groups	Tổng bình phương chêch lệch	Bậc tự do	Phương sai	Tiêu chuẩn kiểm định F
Between groups	SSB = 1648×10^7	df = k-1=3-1=2	$MSB = \frac{SSB}{k-1} = 8241726$	$F = \frac{MSB}{MSW} = 0.156$
Within groups	SSW = 2537×10^{10}	df=N-k=113+181+188-3=479	$MSE = \frac{SSE}{N-k} = 52970224$	

Bảng 5.1. Bảng kết quả mô hinhg ANOVA

Kết luận: Mức ý nghĩa quan sát p − value bằng 0.856 lớn hơn mức ý nghĩa 5% nên chúng ta chưa đủ điều kiện bác bỏ giả thuyết H0.

Kết luận: Không có sự khác biệt về tổng chi phí đặt hàng trung bình của các kho hàng

Hệ số xác định: $R^2 = \frac{SSB}{SSE} \times 100\% = \frac{1648\times10^7}{2537\times10^{10}} \times 100\% = 0.065\%$.

Nhận xét: Với hệ số xác định R^2 là 0.065% trong mô hình ANOVA một yếu tố cho biến order_total giữa ba kho hàng Bakers, Thompson và Nickolson, mô hình cho thấy trong 100% sự biến động của tổng chi phí đặt hàng so với trung bình của nó thì có 0.065% sự khác biệt giữa các kho hàng. Điều này chỉ ra rằng sự khác biệt giữa các kho hàng là không đáng kể.

5.4. Mô hình hồi quy tuyến tính đa bội

Bài toán: Xây dựng mô hình hồi quy tuyến tính đa biến tốt nhất của chi phí vận chuyển delivery_charges theo các biến season, distance_to_nearest_warehouse, is_expedited_delivery, order_price.

Mô hình hồi quy bao gồm:

— Biến phụ thuộc: delivery_charges.

— Biến độc lập: season, distance_to_nearest_warehouse, is_expedited_delivery, order_price.

5.4.1. Xây dựng mô hình hồi quy tuyến tính tốt nhất

Tạo tệp tin gồm các biến delivery_charges, season, order_price, is_expedited_delivery, distance_to_nearest_warehouse để thực hiện mô hình hồi quy tuyến tính đa biến. Sử dụng hàm phân tích ngoại lai và xóa các hàng chứa giá trị ngoại lai và xuất bảng giá trị xem xét các số lượng còn lại của mỗi phần tử sau khi loại bỏ ngoại lai.

```
[1] 472
```

Hình 5.17. Cỡ mẫu của của tệp tin sau khi xử lý ngoại lai

Sau khi tiến hành kiểm tra tính phụ thuộc của các biến độc lập, từ dữ liệu của tệp tin chứa các biến cần để thực hiện mô hình đã được xử lý và loại bỏ ngoại lai, chúng ta tiến hành chia bộ dữ liệu thành 2 phần: train_data (80%) dùng để xây dựng mô hình hồi quy và test_data(20%) dùng để thực hiện dự báo.

Mô hình 1: Sử dụng tất cả các biến độc lập có trong 'linear' để thực hiện hồi quy lai.

- Mô hình được biểu diễn:

delivery_charges = β_0 + β_1 × distance_to_nearest_warehouse + β_2 × seasonSpring + β_3 × seasonSummer + β_4 × seasonWinter + β_5 × order_price + β_6 × is_expedited_deliveryTRUE.

```
Call:
lm(formula = delivery_charges ~ distance_to_nearest_warehouse +
    season + order_price + is_expedited_delivery, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-22.149  -6.973   1.799   5.994  18.410

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.392e+01  1.470e+00  36.668  < 2e-16 ***
distance_to_nearest_warehouse  6.974e+00  8.941e-01   7.800 6.39e-14 ***
seasonSpring                  1.764e+01  1.107e+00  15.931  < 2e-16 ***
seasonSummer                  9.512e+00  1.163e+00   8.179 4.64e-15 ***
seasonWinter                 -1.224e+00  1.165e+00  -1.051    0.294
order_price                  -1.875e-05  5.302e-05  -0.354    0.724
is_expedited_deliveryTrue     1.761e+01  8.196e-01  21.485  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.898 on 370 degrees of freedom
Multiple R-squared:  0.7105,    Adjusted R-squared:  0.7058
F-statistic: 151.3 on 6 and 370 DF,  p-value: < 2.2e-16
```

Hình 5.18. Kết quả của mô hình 1

Nhận xét: Từ kết quả phân tích trên Hình 5.18, ta thu được:

$$\hat{\beta}o = 53.92;\ \hat{\beta}_1 = 6.974;\ \hat{\beta}_2 = 17.64;\ \hat{\beta}_3 = 9.512;$$
$$\hat{\beta}_4 = -1.224;\ \hat{\beta}_5 = -1.875\times 10^{-5};\ \hat{\beta}_6 = 17.61$$

- Như vậy, đường thẳng hồi quy ước lượng theo mô hình 1 là:

delivery_charges = 53.92 + 6.974 × distance_to_nearest_warehouse + 17.64 × seasonSpring + 9.512 × seasonSummer − 1.224×seasonWinter−1.875×10^{-5} × order_price + 17.61 × is_expedited_deliveryTRUE

- Kiểm định các hệ số β_i:

— Giả thuyết H0: $\beta_i = 0$: hệ số βi không có ý nghĩa thống kê.

— Đối thuyết H1: $\beta_i \neq 0$: hệ số βi có ý nghĩa thống kê.

Chúng ta nhận thấy rằng: p-value ứng với các biến: seasonWinter, order-price lớn hơn mức ý nghĩa 5% nên ta chưa đủ điều kiện bác bỏ giả thuyết H0. Vậy hệ số βi ứng với các biến này bằng 0 (không có ý nghĩa thống kê). Tức là các biến này không ảnh hưởng đến delivery_charges. Ta sẽ xem xét loại bỏ các biến này ra khỏi mô hình.

Tuy nhiên, vì biến Season là một biến phân loại với một số phân loại có ý nghĩa, nên việc loại bỏ biến season có thể làm giảm chất lượng của mô hình. Do đó, chúng ta chỉ loại bỏ biến order_price để xây dựng mô hình 2.

Mô hình 2: Bỏ đi biến: order_price từ mô hình 1 lai.

```
Call:
lm(formula = delivery_charges ~ distance_to_nearest_warehouse +
    season + is_expedited_delivery, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-22.030  -7.086   1.768   6.075  18.198

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  53.6740     1.2967  41.394  < 2e-16 ***
distance_to_nearest_warehouse  6.9671     0.8929   7.803 6.21e-14 ***
seasonSpring                 17.6411     1.1058  15.953  < 2e-16 ***
seasonSummer                  9.5198     1.1614   8.197 4.06e-15 ***
seasonWinter                 -1.1998     1.1617  -1.033    0.302
is_expedited_deliveryTrue    17.6020     0.8184  21.508  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.889 on 371 degrees of freedom
Multiple R-squared:  0.7104,    Adjusted R-squared:  0.7065
F-statistic:   182 on 5 and 371 DF,  p-value: < 2.2e-16
```

Hình 5.19. Kết quả của mô hình 2

Nhận xét: Từ kết quả của mô hình 2 trên Hình 5.19, đường thẳng hồi quy ước lượng theo mô hình 2 là:

delivery_charges = 53.*674* + 6.*9671* × distance_to_nearest_warehouse + 17.*6411* × seasonSpring + 9.*5198* × seasonSummer − 1.1998 × seasonWinter + 17.*602* × is_expedited_deliveryTRUE.

5.4.2. Đánh giá mô hình 2:

a. Đánh giá các biến trong mô hình

— *Intercept*: Hệ số ước lượng là 53.674, nghĩa là khi tất cả các biến độc lập đều bằng 0, giá trị trung bình của *delivery_charges* sẽ là 53.674.

— *Distance_to_nearest_warehouse*: Hệ số ước lượng là 6.9671 với p-value rất nhỏ, điều này cho thấy rằng khoảng cách tới nhà kho gần nhất có ảnh hưởng đáng kể đến *delivery_charges*, khi khoảng cách tăng 1 đơn vị, giá trị của biến phụ thuộc tăng 6.9671 đơn vị.

— *SeasonSpring*: Hệ số ước lượng là 17.6411 với p-value rất nhỏ, điều này cho thấy rằng mùa Xuân (so với mùa Thu, là biến tham chiếu) có ảnh hưởng đáng kể và dương đến *delivery_charges*, tức là khi giao hàng vào mùa Xuân, chi phí giao hàng trung bình tăng thêm 17.6411 đơn vị.

— *SeasonSummer*: Hệ số ước lượng là 9.5198 với p-value rất nhỏ, cho thấy mùa Hè (so với mùa Thu, là biến tham chiếu) cũng có ảnh hưởng đáng kể đến *delivery_charges* nhưng mức tăng ít hơn so với mùa Xuân. Khi giao hàng vào mùa hè, chi phí giao hàng trung bình tăng thêm 9.5198 đơn vị.

— *SeasonWinter*: Hệ số ước lượng là -1.1998 với p-value = 0.302, lớn hơn mức ý nghĩa 5%, nên biến này không có ý nghĩa thống kê trong mô hình. Điều này có

nghĩa là mùa Đông không có ảnh hưởng đáng kể đến *delivery_charges* so với mùa Thu.

— *is_expedited_deliveryTrue*: Hệ số ước lượng là 17.602 với p-value rất nhỏ, cho thấy việc chọn giao hàng nhanh (*expedited delivery*) có ảnh hưởng đáng kể đến chi phí giao hàng, tức là khi chuyển từ không giao hàng nhanh sang giao hàng nhanh thì *delivery_charges* sẽ tăng thêm 17.602 đơn vị.

b. Đánh giá mức độ phù hợp của mô hình

— Giá trị *Multiple R-squared* (hệ số xác định R^2) là 0.7104, giải thích rằng trong 100% sự biến động của *delivery_charges* so với trung bình của nó, có 71.04% là do các biến trong mô hình gây ra.

— Giá trị *Adjusted R-squared* (hệ số hiệu chỉnh R^2) là 0.7065, dựa trên số lượng biến trong mô hình. Nó thường thấp hơn hệ số xác định do loại ra các biến không cần thiết trong mô hình. Với giá trị 0.7065 nghĩa là so với trung bình của nó thì có 70.65% sự biến động của *delivery_charges* là do các biến trong mô hình gây ra.

— **Ưu điểm**: Đây là một mức R^2 khá tốt, cho thấy các biến độc lập có liên quan khá chặt chẽ đến biến phụ thuộc.

— **Nhược điểm**: Tuy nhiên còn 29.35% sự biến động không được giải thích, cho thấy rằng những yếu tố khác không nằm trong mô hình cũng đang ảnh hưởng đến chi phí giao hàng.

— Giá trị *Adjusted R-squared* là 0.7065 dựa trên số lượng biến trong mô hình rất gần với R^2 là 0.7104, cho thấy các biến trong mô hình là hợp lý và không có biến dư thừa.

— Giá trị *Residual standard error* là 7.889 tương đối nhỏ so với giá trị trung bình của biến phụ thuộc, điều này cho thấy mô hình có độ chính xác tương đối tốt trong việc dự đoán chi phí giao hàng.

c. Kiểm định sự phù hợp của hàm hồi quy tuyến tính đa biến

— Giả thuyết H0: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_i = 0$ (mô hình không phù hợp).

— Giả thuyết H1: Tồn tại β_i khác 0 (mô hình phù hợp).

— **Kiểm định bằng p-value**: p-value nhỏ hơn rất nhiều so với mức ý nghĩa 5% nên ta đủ điều kiện bác bỏ giả thuyết H0. Vậy nghĩa là có ít nhất một trong các biến độc lập *βi* khác 0 ảnh hưởng đáng kể đến chi phí giao hàng, tức là mô hình có ý nghĩa thống kê.

— **Kiểm định bằng F-statistic**:

+ Ta có: $RR = (f_{\alpha,k,N-k-1}, +\infty) = (f_{0.05,5,371}, +\infty) = (2.25; +\infty)$.
+ $F = 182$.

— Vì giá trị của F-statistic (182) rất cao và thuộc miền bác bỏ nên ta bác bỏ giả thuyết H0.

Vậy nghĩa là có ít nhất một trong các biến độc lập βi khác 0 ảnh hưởng đáng kể đến chi phí giao hàng, tức là mô hình có ý nghĩa thống kê.

5.4.3. Các giả định của mô hình hồi quy

$Y_i = \beta_0 + \beta_1.X_1 + \ldots..\beta_i,\ i = 1, \ldots\ldots n$

— Tính tuyến tính của dữ liệu: mối quan hệ giữa biến dự báo X và biến phụ thuộc Y
được giả sử là tuyến tính.

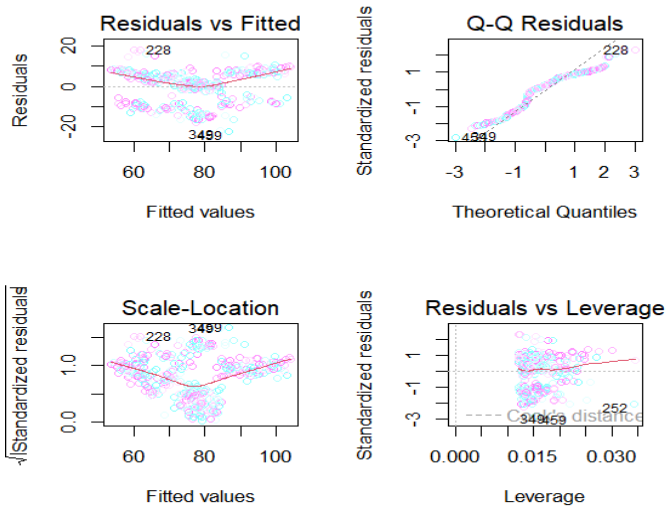
— Sai số có kỳ vọng bằng 0.

— Phương sai của các sai số là hằng số.

— Sai số có phân phối chuẩn. $\varepsilon \sim N(0,\sigma^2)$

— Các sai số $\varepsilon_1, \ldots, \varepsilon_n$ thì độc lập với nhau.

Vẽ đồ thị: thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình.



Hình 5.20. Đồ thị thặng dư

Nhận xét: Từ các đồ thị trên Hình 5.20 ta có các nhận xét sau:

Đồ thị 1: Vẽ các sai số hồi quy tương ứng với giá trị dự báo, dùng để kiểm tra giả định:

— Y và các biến độc lập X có quan hệ tuyến tính.

— Sai số có kỳ vọng bằng 0.

— Phương sai các sai số là hằng số.

Dựa vào kết quả cho thấy:

— Đường màu đỏ không là đường thẳng nằm ngang nên giả định Y và các biến độc lập X có quan hệ tuyến tính không thỏa mãn.

— Đường màu đỏ không nằm sát đường $y = 0$ nên giả định sai số có kỳ vọng bằng 0 thỏa mãn.

— Các giá trị sai số chưa phân tán ngẫu nhiên dọc theo đường màu đỏ, nên giả định phương sai các sai số là hằng số chưa thỏa mãn.

Đồ thị 2: Vẽ các sai số hồi quy được chuẩn hoá, dùng để kiểm tra giả định sai số có phân phối chuẩn. Dựa vào kết quả cho thấy: các quan trắc không nằm trên đường thẳng kỳ vọng phân phối chuẩn nên giả định sai số có phân phối chuẩn không thỏa mãn.

Đồ thị 3: Vẽ căn bậc 2 các sai số được chuẩn hoá, dùng để kiểm tra giả định phương sai các sai số là hằng số. Dựa vào kết quả cho thấy: Các giá trị sai số chưa phân tán ngẫu nhiên dọc theo đường màu đỏ, nên giả định phương sai các sai số là hằng số chưa thực sự thỏa mãn.

Đồ thị 4: Vẽ ra những điểm có gây ảnh hưởng cao trong bộ dữ liệu, cụ thể là các điểm 349, 459, 252. Tuy nhiên, không có điểm nào vượt ra khỏi đường Cook's distance. Vì vậy, ta không cần phải loại bỏ các điểm ra khỏi bộ dữ liệu.

5.4.4. Dự đoán

	predicted_value	delivery_charges
3	61.30227	65.87
7	76.02967	74.84
11	78.38025	78.25
16	81.29599	80.69
21	79.12934	63.51
29	76.94224	77.24
32	68.14528	70.09
33	87.05501	92.94
52	84.88057	89.45
58	78.96758	65.32

Hình 5.21. Kết quả dự đoán

Nhận xét: Từ kết quả dự đoán trên Hình 5.21, chúng ta có thể thấy rằng không có sự sai khác quá lớn giữa giá trị dự đoán và giá trị thực tế của delivery_charges, nên phương trình hồi quy tuyến tính ở mô hình 2 là khá phù hợp.

6. NGUỒN DỮ LIỆU VÀ CODE

Link dữ liệu: Tại đây

7. TÀI LIỆU THAM KHẢO

1. *Minh Lan (16/09/2019), "Biểu đồ tần suất Histogram là gì? Trình tự lập biểu đồ tần suất". Truy cập từ:* https://tinyurl.com/yjkbcun4
2. *Nguyễn Văn Tuấn (Garvan Institute of Medical Research Sydney, Australia), "Phân tích số liệu và biểu đồ bằng R". Truy cập từ:* https://tinyurl.com/yc67nycc
3. *Kiểm định phân phối chuẩn với R (normality test with R) | Phân Tích Thống Kê 7 |Learn to do SCIENCE. Truy cập từ:* https://www.youtube.com/watch?v=hioS2jpoVtk