MAS Workshop

Department of Statistics - USJ

2024-04-02

```r
# install packages
install.packages("readxl", "dplyr", "ggplot2", "magrittr", "car", "DescTools",
                 "tidyr","janitor")
```

```r
# load the packages

library(readxl)    # to load the data set
library(dplyr)     # select operator
library(ggplot2)   # to create plots
library(magrittr)  # pipe operator
library(car)       # to obtain vif value
library(DescTools) # to obtain the mode
library(tidyr)
library(janitor)
```

Descriptive Statistics

Load the data

```r
data_descriptive <- read_xlsx("Descriptive Statistics Data.xlsx")
```

Glimpse on the dataset

```r
glimpse(data_descriptive)
```

```
Rows: 12,947
Columns: 9
$ `Shipping Type`  <chr> "Courier", "Courier", "Courier", "Courier", "Courier"~
$ SMV              <dbl> 4.880, 4.880, 4.734, 4.734, 4.734, 4.734, 4.880, 4.88~
$ Plant            <chr> "D051", "D051", "D051", "D051", "D051", "D051", "D051~
$ `Order Qty`      <dbl> 200, 200, 200, 200, 200, 200, 3, 197, 3, 197, 3, 197,~
$ Earnings         <dbl> 509.2671800, 836.6231200, 812.6663206, 812.6663206, 8~
$ Date             <chr> "2023-06-01", "2023-06-01", "2023-06-01", "2023-06-01~
$ `Customer Group` <chr> "Abercrombie & Fitch", "Abercrombie & Fitch", "Abercr~
$ Earn             <dbl> 219.814988, 235.986348, 332.006986, 332.006986, 311.5~
$ EPH              <dbl> 45.12728, 48.44721, 25.15586, 25.15586, 38.25048, 38.~
```

slide number: 43

one way frequency table

```r
table(data_descriptive$`Shipping Type`)
```

```
Air Collect     Courier Sea Collect 
         87          86         937 
```

```r
tabyl(data_descriptive$`Shipping Type`,sort=TRUE,show_na = FALSE)
```

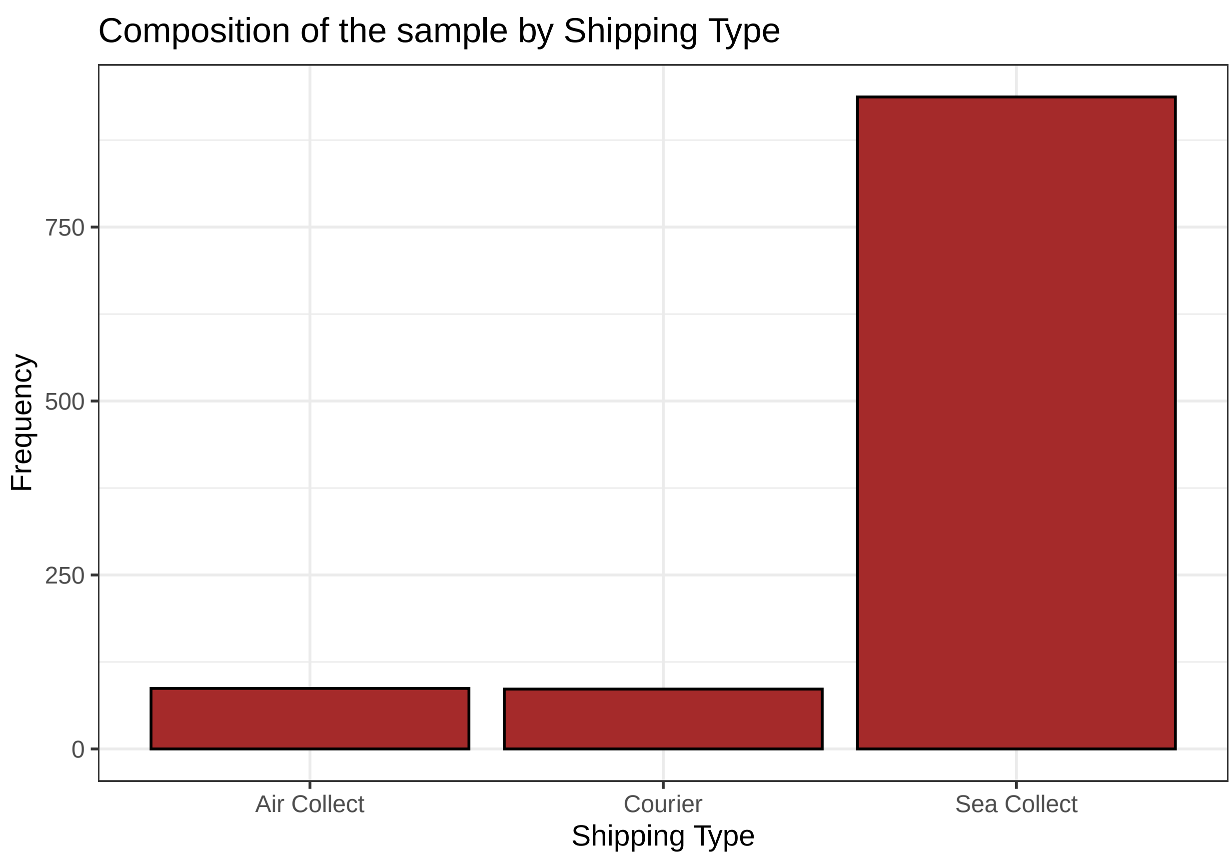
```
 data_descriptive$`Shipping Type`   n    percent
                      Air Collect  87 0.07837838
                          Courier  86 0.07747748
                      Sea Collect 937 0.84414414
```

```r
# to obtain tidy output
```

slide number: 44

barchart

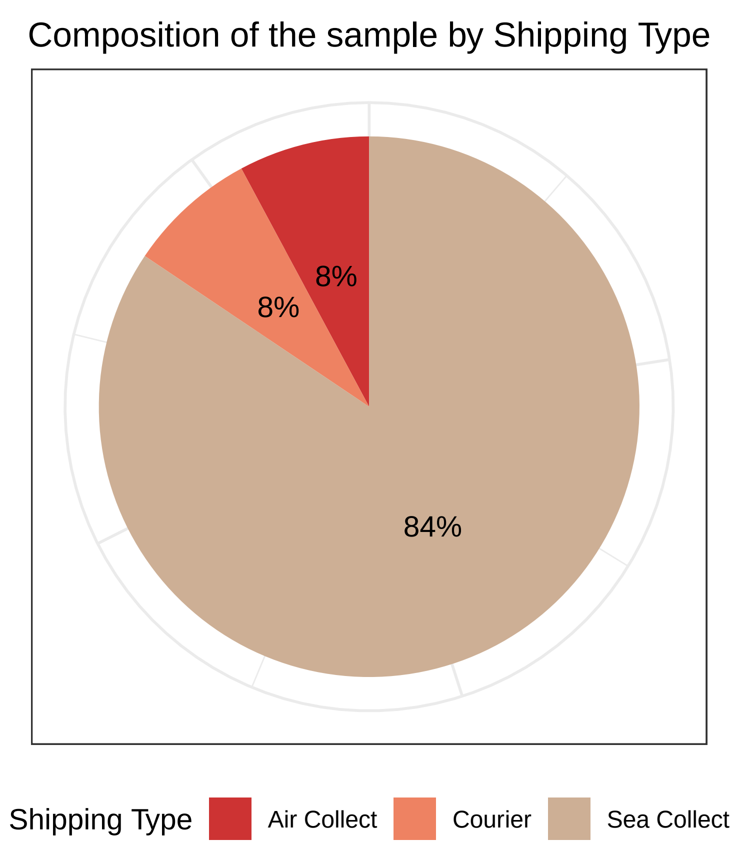
```
data_descriptive %>%
  select(`Shipping Type`) %>%
  na.omit() %>%
  ggplot(aes(x = as.factor(`Shipping Type`))) +
  geom_bar(color="black",
           fill="brown" ) +
  theme_bw() +
  labs(title = "Composition of the sample by Shipping Type",
       x = "Shipping Type",
       y = "Frequency")
```



slide number: 45

pie chart

```
data.frame(Shipping_Type = c("Air Collect", "Courier", "Sea Collect"),
           Frequency = c(87, 86, 937)) %>%
  ggplot(aes(x = "", y = Frequency,
             fill = Shipping_Type)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  geom_text(aes(label = paste0(
    round((Frequency/sum(Frequency))*100), "%")),
    position = position_stack(vjust = 0.5)) +
  theme_bw() +
  scale_fill_manual(values=c("brown3", "salmon2", "peachpuff3")) +
  labs(x = NULL, y = NULL,
       fill = "Shipping Type",
       title = "Composition of the sample by Shipping Type") +
  theme(axis.line = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  theme(legend.position = "bottom")
```



slide number: 46

summary measures

```r
# mean
mean(data_descriptive$SMV,
     na.rm = TRUE)
```

```
[1] 10.39356
```

```r
# median
median(data_descriptive$SMV,
       na.rm = TRUE)
```

```
[1] 9.741
```

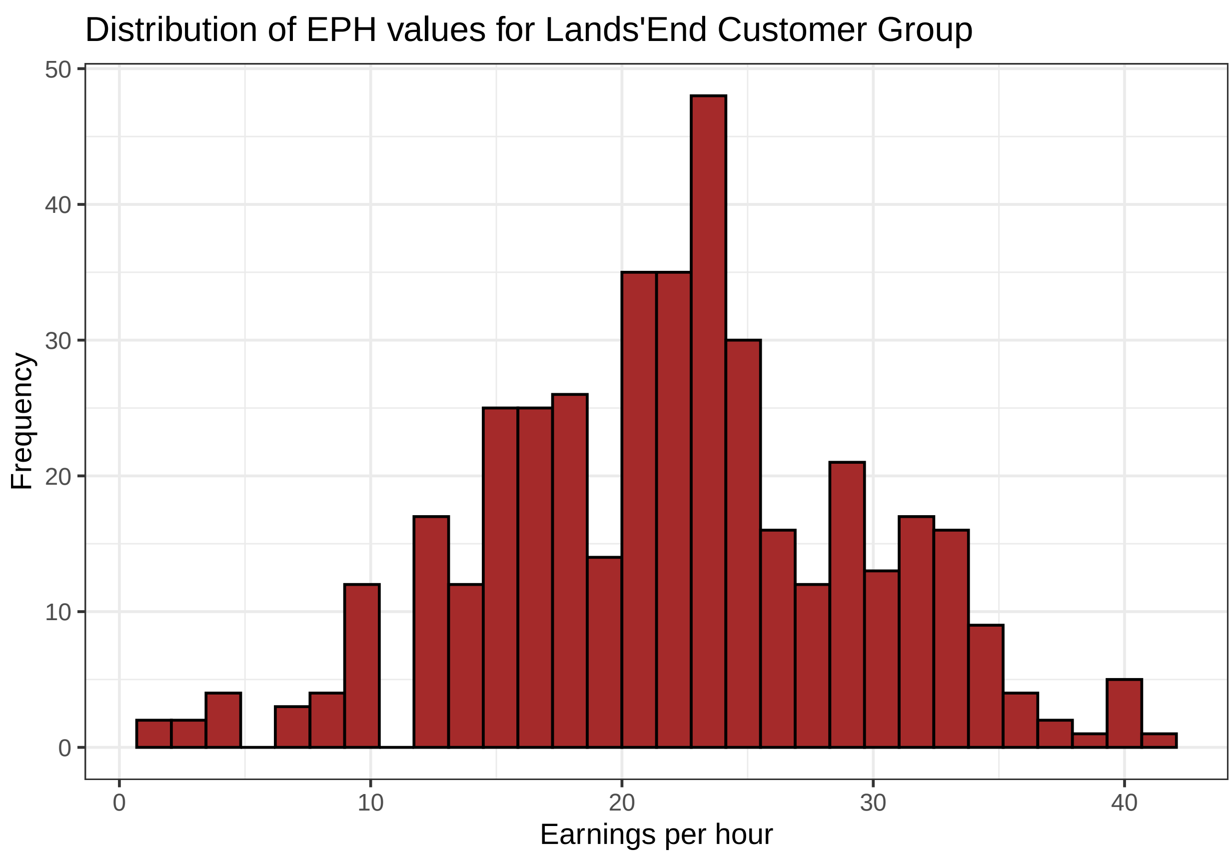
```r
# mode
Mode(data_descriptive$SMV,
     na.rm = TRUE)
```

```
[1] 14.088
attr(,"freq")
[1] 71
```

slide number: 47

histogram - symmetric

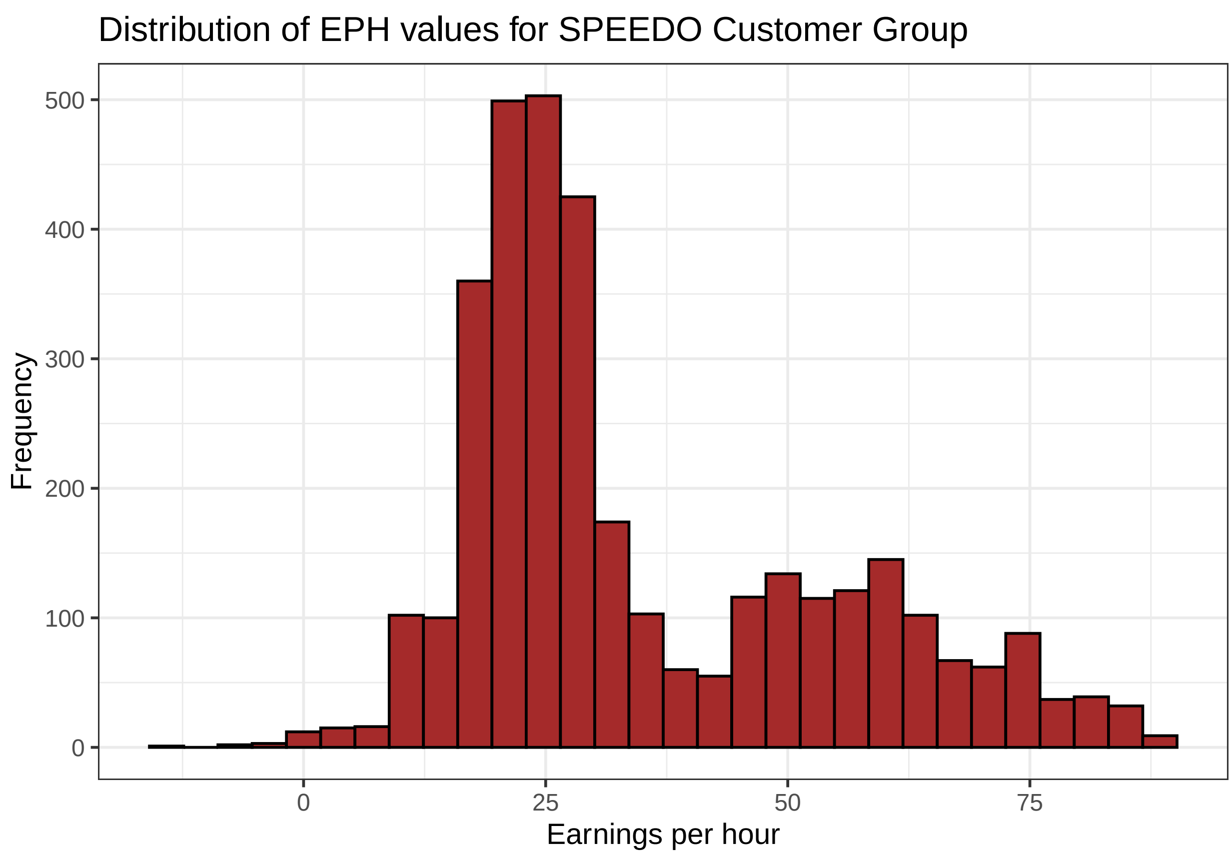
```r
# preparing data
symmetric <- data_descriptive %>% filter(`Customer Group` == "Lands'End")
outliers <- boxplot(symmetric$EPH, plot=FALSE)$out
symmetric <- symmetric[-which(symmetric$EPH %in% outliers), ]

# histogram
symmetric %>%
  select(EPH) %>%
  ggplot(aes(x = EPH)) +
  geom_histogram(color = "black",
                 fill = "brown") +
  theme_bw() +
  labs(title = "Distribution of EPH values for Lands'End Customer Group",
       x = "Earnings per hour",
       y = "Frequency")
```



slide number: 48

histogram - positively skewed

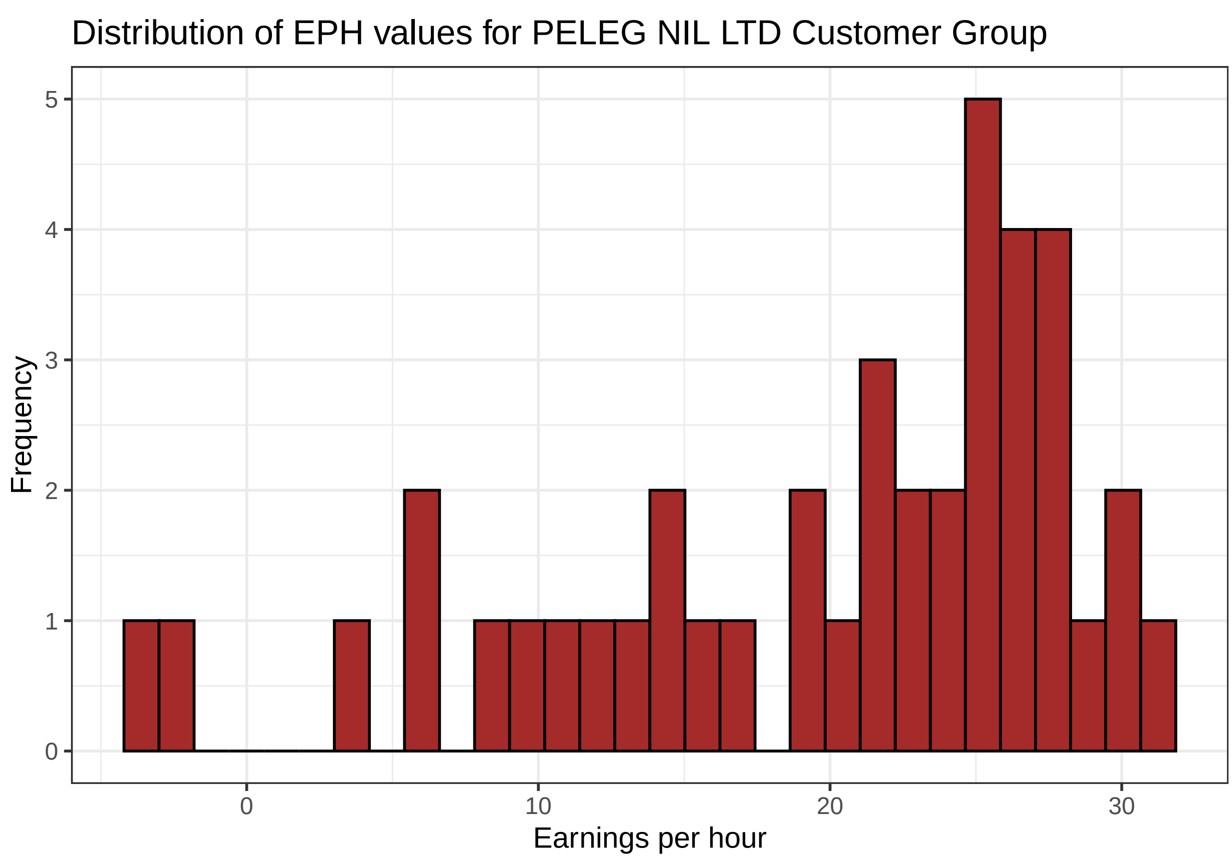
```
# preparing data
pos.skewed <- data_descriptive %>% filter(`Customer Group` == "SPEEDO NA")
outliers <- boxplot(pos.skewed$EPH, plot=FALSE)$out
pos.skewed <- pos.skewed[-which(pos.skewed$EPH %in% outliers), ]

#histogram
pos.skewed %>%
  select(EPH) %>%
  ggplot(aes(x = EPH)) +
  geom_histogram(color = "black",
                 fill = "brown") +
  theme_bw() +
  labs(title = "Distribution of EPH values for SPEEDO Customer Group",
       x = "Earnings per hour",
       y = "Frequency")
```



slide number: 49

histogram - negatively skewed

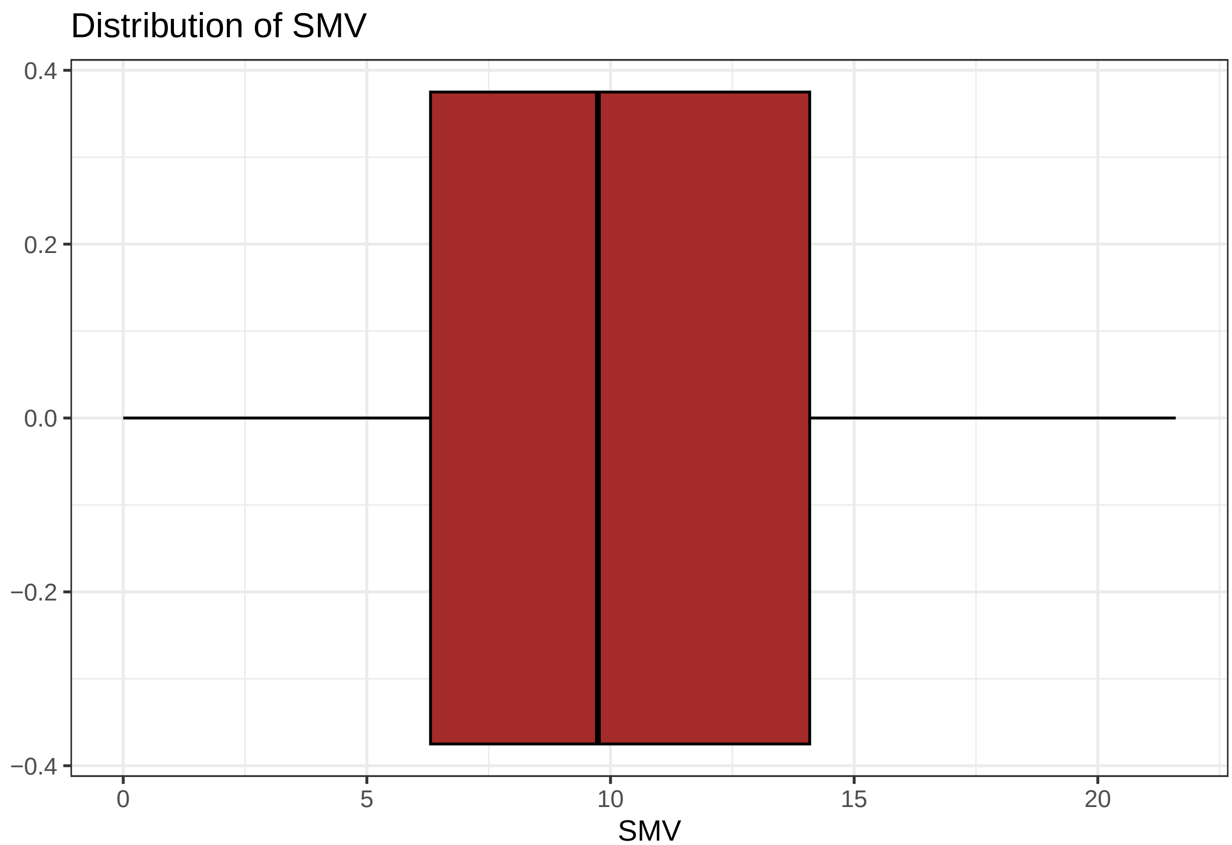
```
# preparing data
neg.skewed <- data_descriptive %>% filter(`Customer Group` == "PELEG NIL LTD")

#histogram
neg.skewed %>%
  select(EPH) %>%
  ggplot(aes(x = EPH)) +
  geom_histogram(color = "black",
                 fill = "brown") +
  theme_bw() +
  labs(title = "Distribution of EPH values for PELEG NIL LTD Customer Group",
       x = "Earnings per hour",
       y = "Frequency")
```



slide number: 50

boxplot

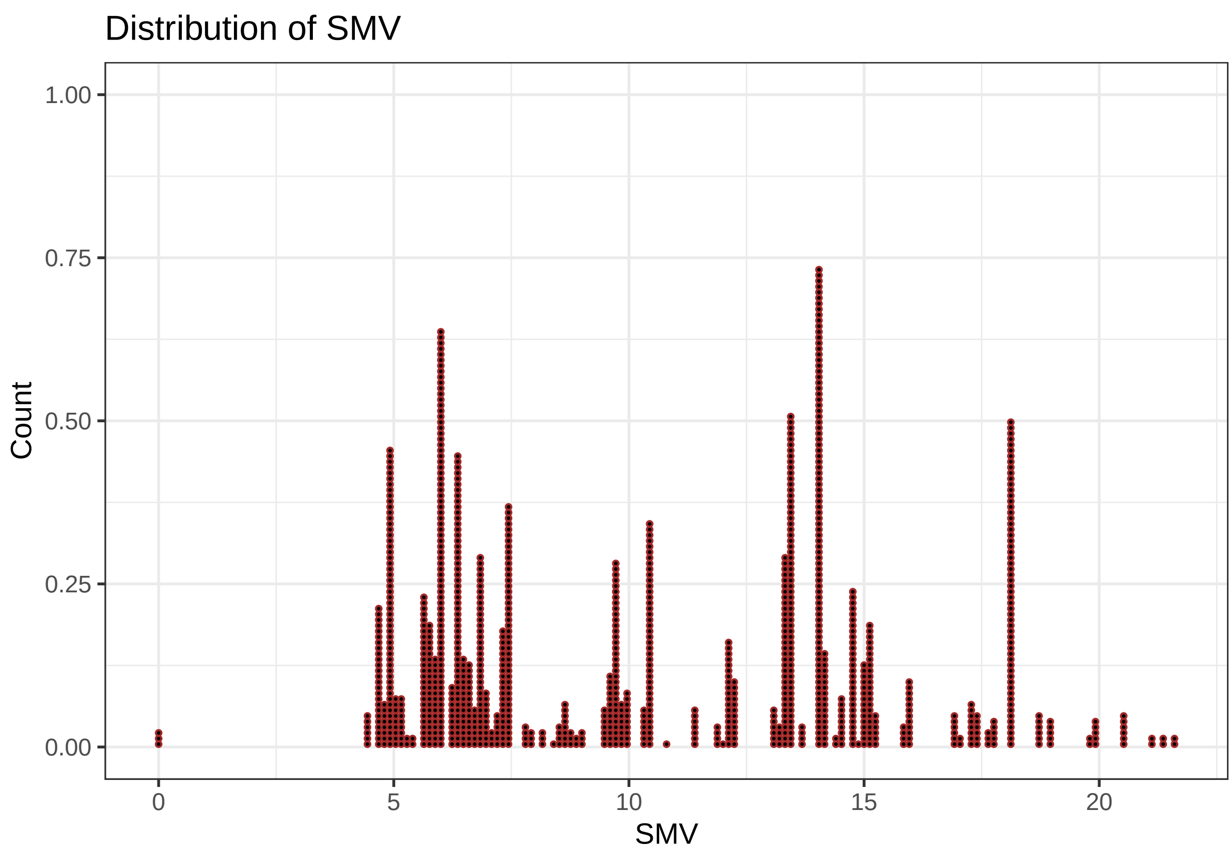
```
data_descriptive %>%
  select(SMV) %>%
  na.omit() %>%
  ggplot(aes(x = SMV)) +
  geom_boxplot(color = "black",
               fill = "brown" ) +
  theme_bw() +
  labs(title = "Distribution of SMV")
```



slide number: 51

dot plot

```
data_descriptive %>%
ggplot(aes(x = SMV)) +
geom_dotplot(method="histodot", binwidth = 0.12,col = "brown") +
labs(x = "SMV", y = "Count", title = "Distribution of SMV") +
theme_bw()
```



slide number: 54

two way frequency table

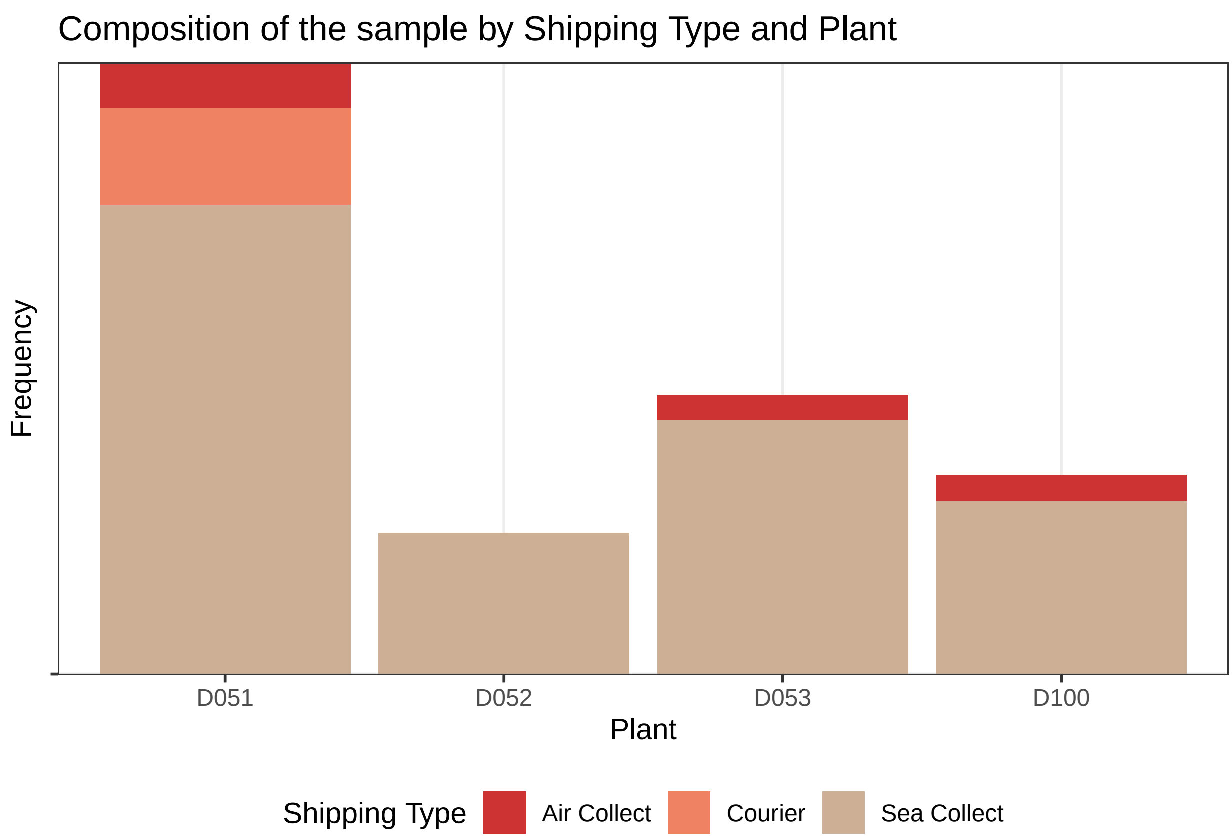
```
table(data_descriptive$`Shipping Type`,
      data_descriptive$Plant)
```

```
              D051 D052 D053 D100
  Air Collect   40    0   24   23
  Courier       86    0    0    0
  Sea Collect  422  126  235  154
```

slide number: 55

stacked bar chart

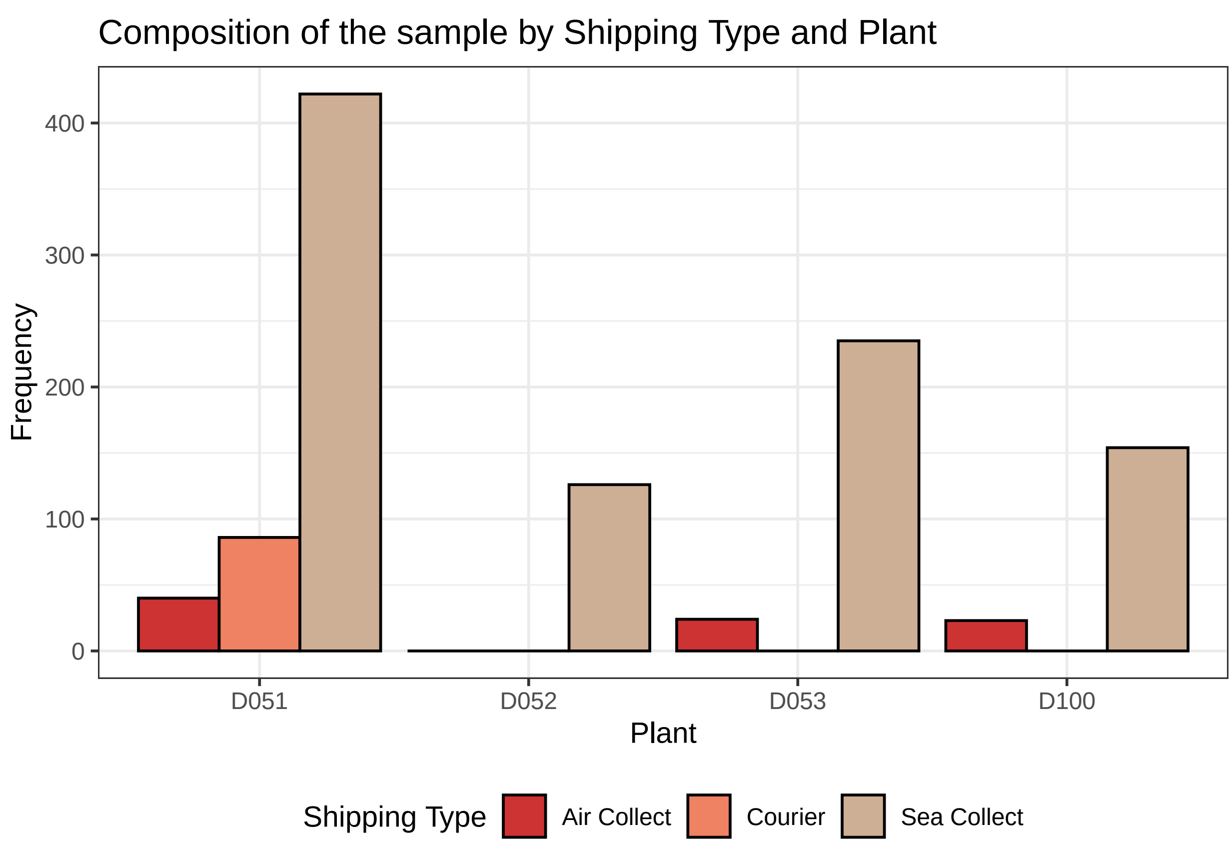
```
data_descriptive %>%
  na.omit() %>%
  ggplot(aes(x = Plant, y = "", fill = `Shipping Type`)) +
  geom_bar(stat = "identity") +
  theme_bw() +
  scale_fill_manual(values = c("brown3", "salmon2", "peachpuff3")) +
  labs(x = "Plant",
       y = "Frequency",
       fill = "Shipping Type",
       title = "Composition of the sample by Shipping Type and Plant") +
  theme(legend.position = "bottom")
```



slide number: 56

cluster bar chart

```
table(data_descriptive$`Shipping Type`,
      data_descriptive$Plant) %>%
  as.data.frame() %>%
  ggplot(aes(x = Var2, y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", position = position_dodge(), col = "black") +
  theme_bw() +
  scale_fill_manual(values=c("brown3", "salmon2", "peachpuff3")) +
  labs(x = "Plant",
       y = "Frequency",
       fill = "Shipping Type",
       title = "Composition of the sample by Shipping Type and Plant") +
  theme(legend.position = "bottom")
```



slide number: 57

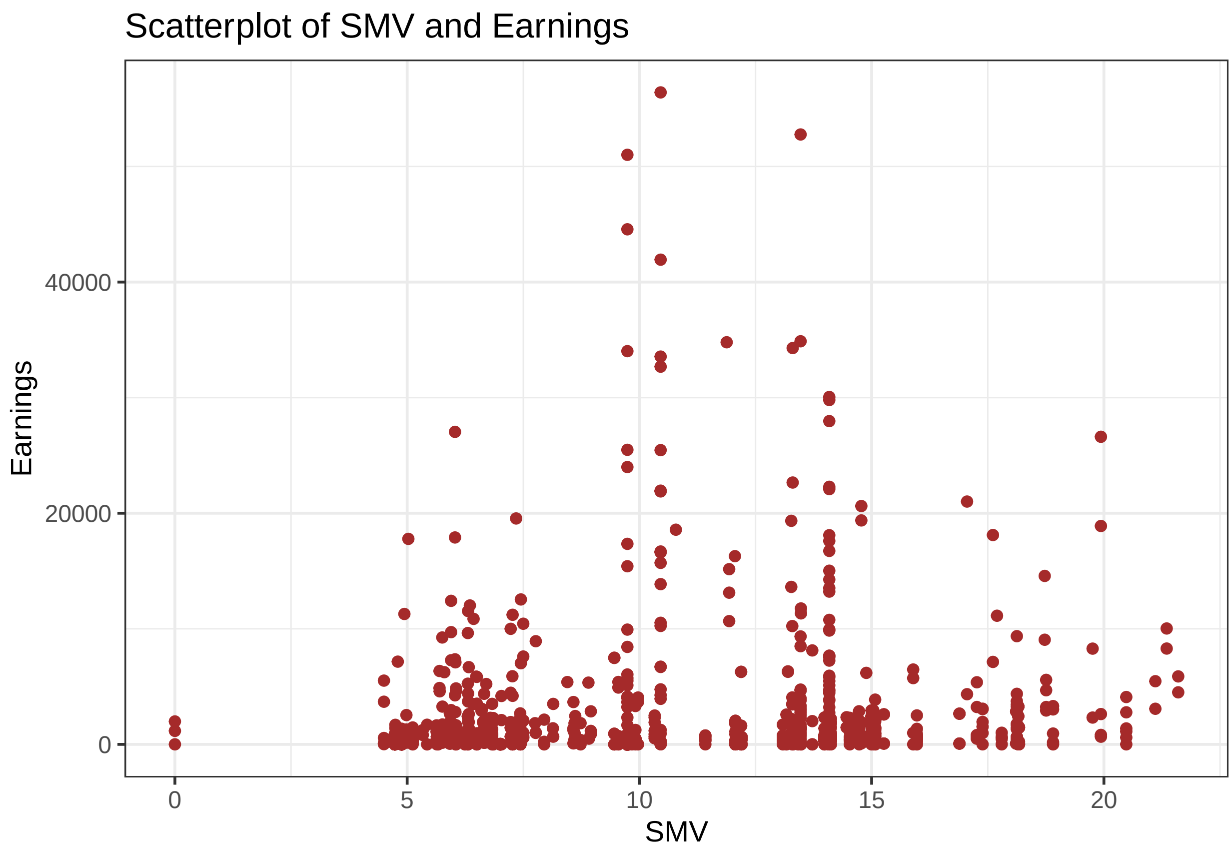
scatterplot - positive linear

```
data_descriptive %>%
  ggplot(aes(x = `Order Qty`,
             y = Earnings)) +
  geom_point(col = "brown") +
  theme_bw() +
  labs(title = "Scatterplot of Order Quantity and Earnings",
       x = "Order Quantity",
       y = "Earnings")
```

slide number: 58

scatterplot - no linear

```
data_descriptive %>%
  na.omit() %>%
  ggplot(aes(x = SMV,
             y = Earnings))+
  geom_point(col = "brown") +
  theme_bw() +
  labs(title = "Scatterplot of SMV and Earnings",
       x = "SMV",
       y = "Earnings")
```

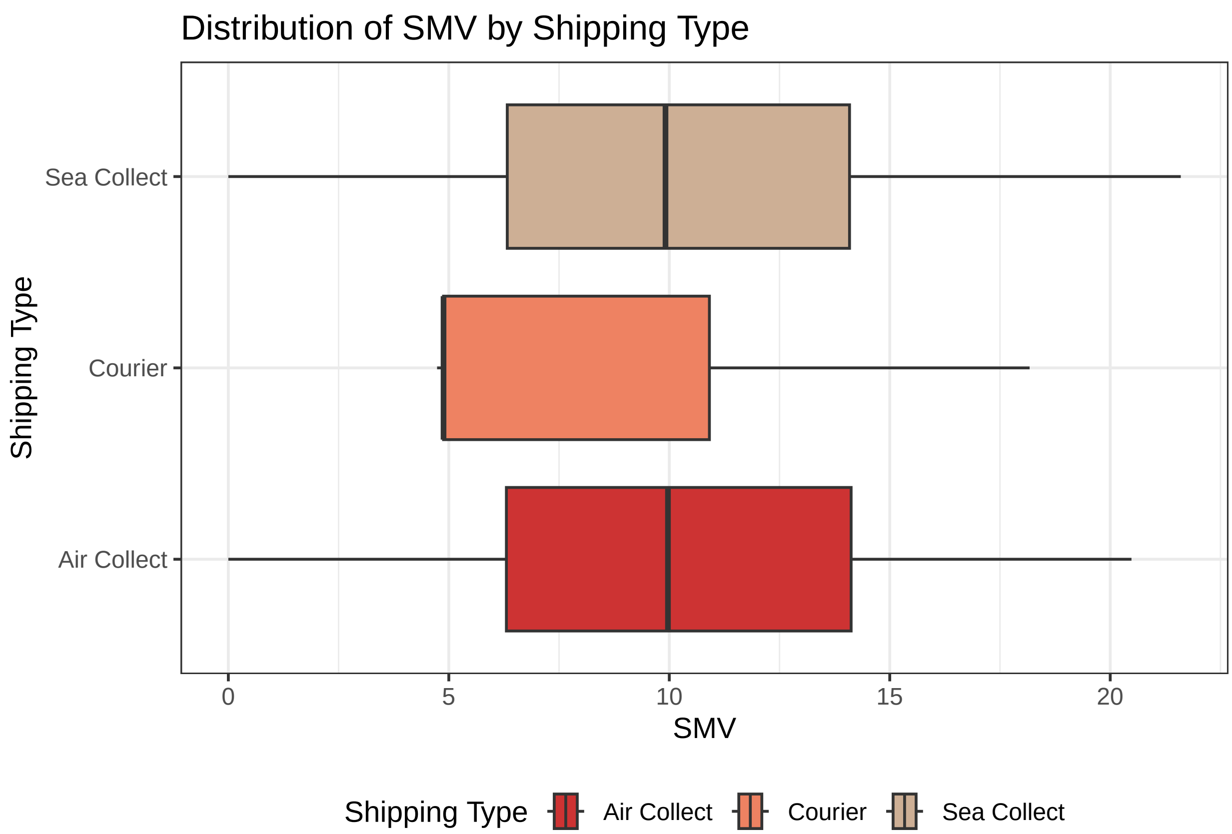


slide number: 59

boxplot with groups

```
data_descriptive %>%
  select(SMV,
```

```
         `Shipping Type`) %>%
  na.omit() %>%
  ggplot(aes(x = SMV,
             y = `Shipping Type`,
             fill = `Shipping Type`)) +
  geom_boxplot() +
  theme_bw() +
  scale_fill_manual(values=c("brown3", "salmon2", "peachpuff3")) +
  labs(x = "SMV",
       y = "Shipping Type",
       fill = "Shipping Type",
       title = "Distribution of SMV by Shipping Type") +
  theme(legend.position = "bottom")
```

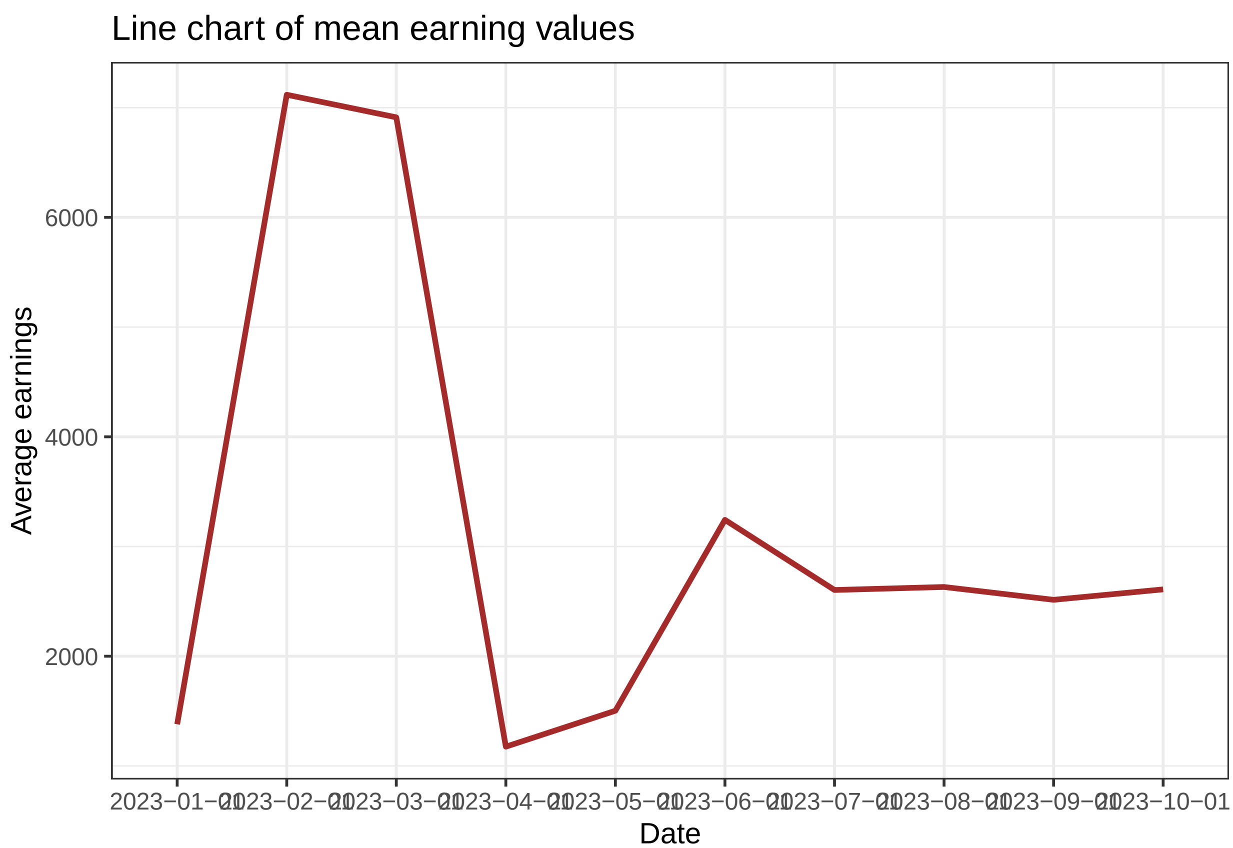


slide number: 61

line chart

```
# preparing data
line_data <- data_descriptive %>% group_by(Date) %>%
  summarise(Mean.Earnings = mean(Earn)) %>% drop_na()
```

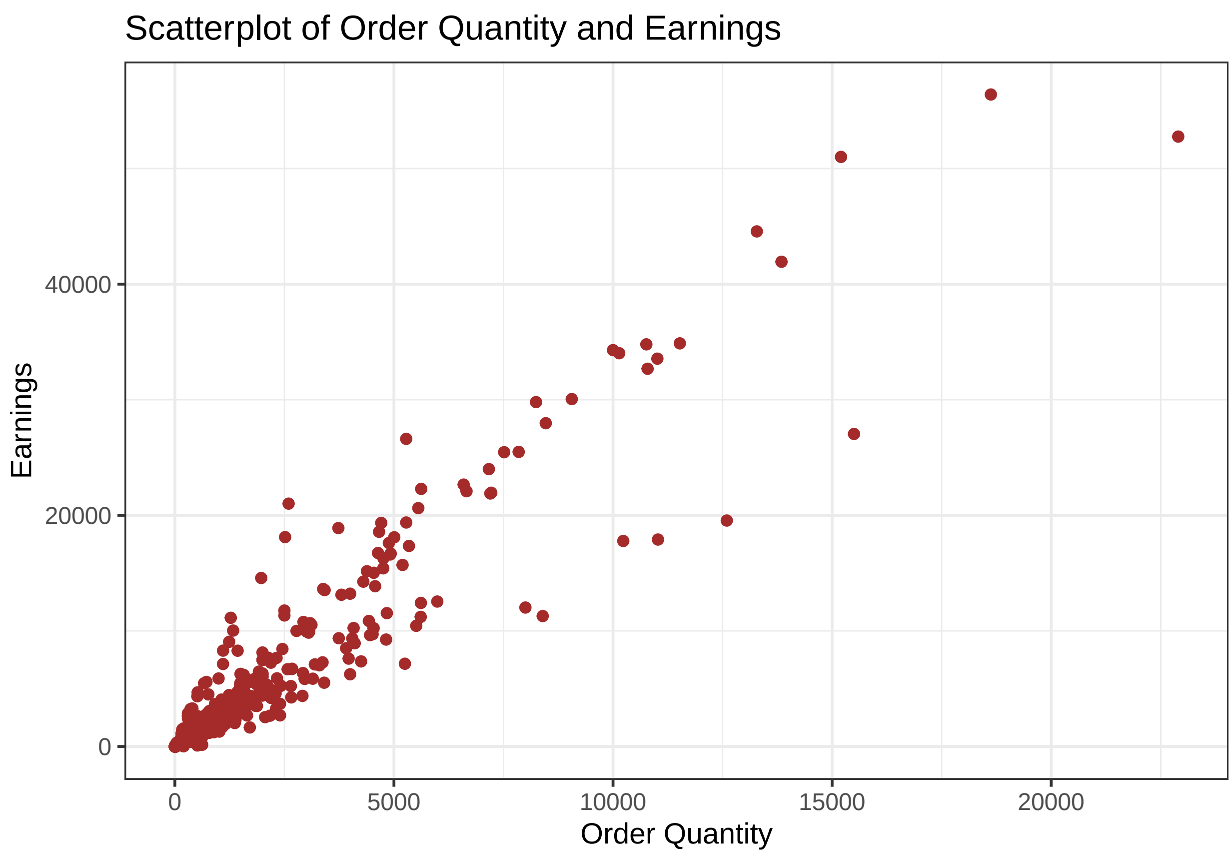
```r
# line chart
line_data %>%
  ggplot(aes(x = Date, y = Mean.Earnings, group=1)) +
  geom_line(color="brown", size=1) +
  theme_bw() +
  labs(title = "Line chart of mean earning values",
       x = "Date",
       y = "Average earnings")
```



Correlation Analysis

slide number: 65

```r
# scatterplot - positive linear
data_descriptive %>%
  na.omit() %>%
  ggplot(aes(x = `Order Qty`,
             y = Earnings)) +
  geom_point(col = "brown") +
  theme_bw() +
  labs(title = "Scatterplot of Order Quantity and Earnings",
       x = "Order Quantity",
       y = "Earnings")
```



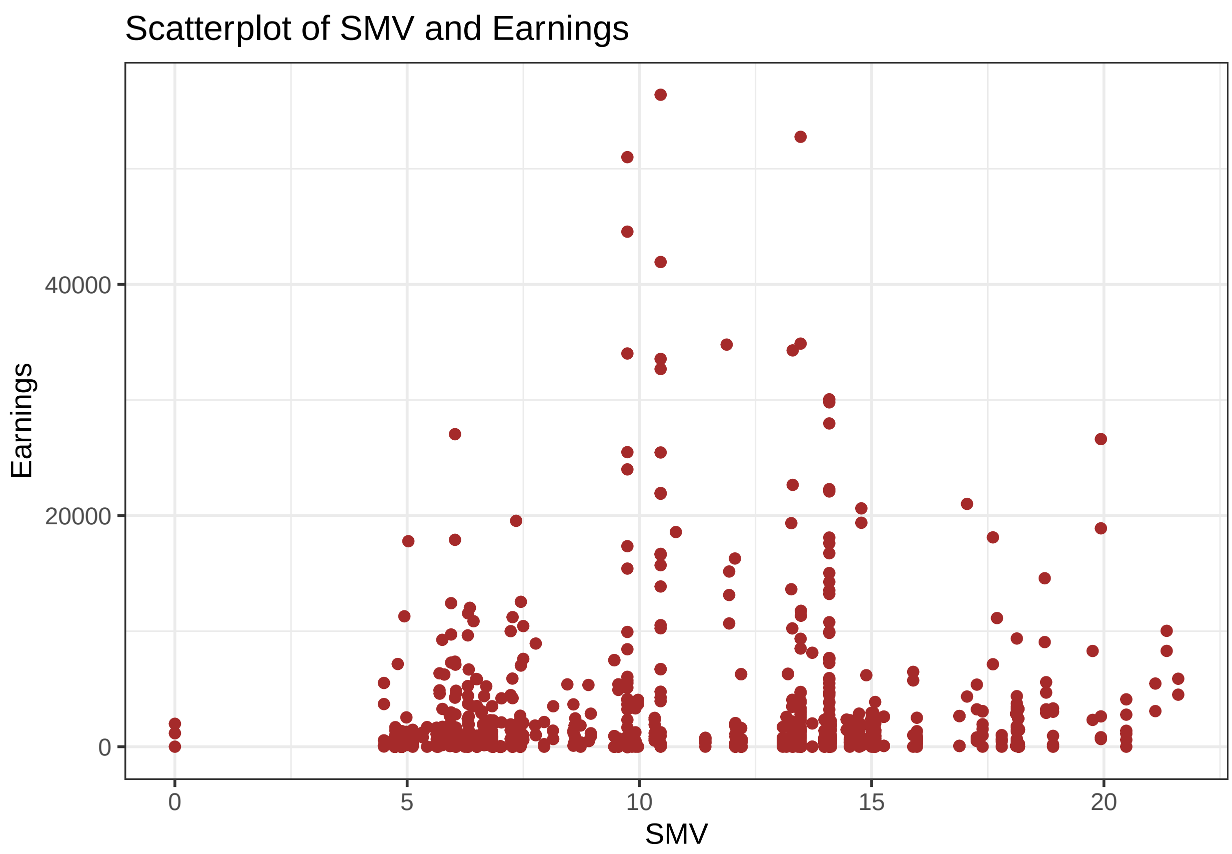

```r
# positive linear relationship

# correlation value
cor(x = data_descriptive$`Order Qty`,
    y = data_descriptive$Earnings)
```

```
[1] NA
```

```
# r = 0.9371457
```

slide number: 66

```
# scatterplot - no linear
data_descriptive %>%
  na.omit() %>%
  ggplot(aes(x = SMV,
             y = Earnings))+
  geom_point(col = "brown") +
  theme_bw() +
  labs(title = "Scatterplot of SMV and Earnings",
       x = "SMV",
       y = "Earnings")
```




```
# no linear relationship

# correlation value
cor(data_descriptive$SMV,
    data_descriptive$Earnings,
    use = "complete.obs")
```

```
[1] 0.1142413
```

```
# 0.1142413
```

Hypothesis Testing

One sample test for mean - Slide no 73

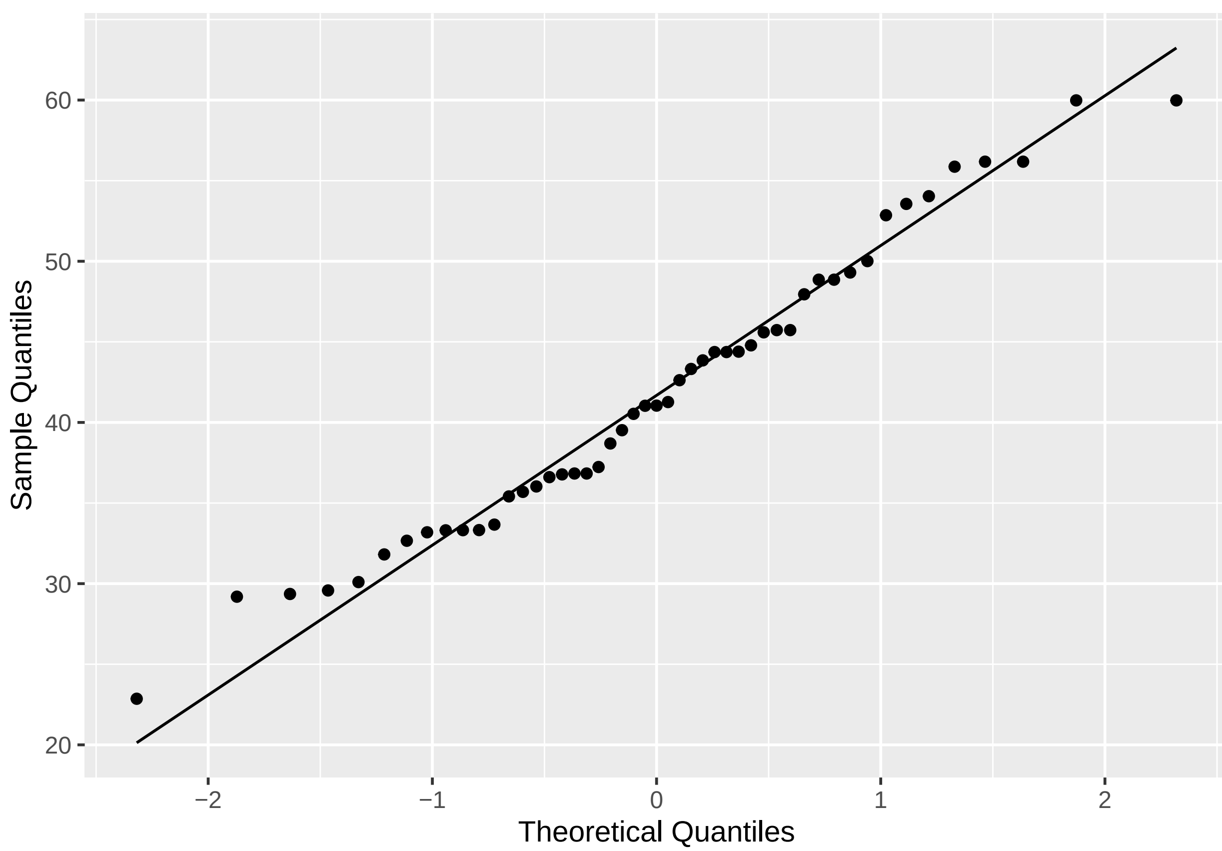
Example: Suppose we want to test whether the mean earnings per hour for Outer Known customer group is less than 50 at 5% significance level.

```
# loading dataset
Hypothesis.data <- read_excel("Hypothesis Data.xlsx")
```

Step 1: Check whether Earnings per hour values are normally distributed

Normal probability plot

```
ggplot(Hypothesis.data, aes(sample = Earnings.per.hour)) + stat_qq() +
  stat_qq_line() +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles")
```



Normality test

```
shapiro.test(Hypothesis.data$Earnings.per.hour)
```

```
	Shapiro-Wilk normality test
```

```
data:  Hypothesis.data$Earnings.per.hour
W = 0.97508, p-value = 0.3806
```

Hypothesis to be tested:

H0: Data are normally distributed.

H1: Data are not normally distributed.

According to the Shapiro-Wilk normality test p-value = 0.3806 > 0.05.

Hence, We can conclude that Earnings per hour values are normally distributed.

Step 2: Perform the t-test

```
t.test(Hypothesis.data$Earnings.per.hour, alternative = "less", mu = 50)
```

```
    One Sample t-test

data:  Hypothesis.data$Earnings.per.hour
t = -6.5365, df = 48, p-value = 1.89e-08
alternative hypothesis: true mean is less than 50
95 percent confidence interval:
     -Inf 43.84388
sample estimates:
mean of x
 41.71904
```

Since p-value = 1.89e-08 < 0.05, we reject null hypothesis.

Hence, there is sufficient evidence to suggest that the mean earnings per hour for the Outer Known customer group is less than 50.

Two sample test for comparison between means - Slide no 75

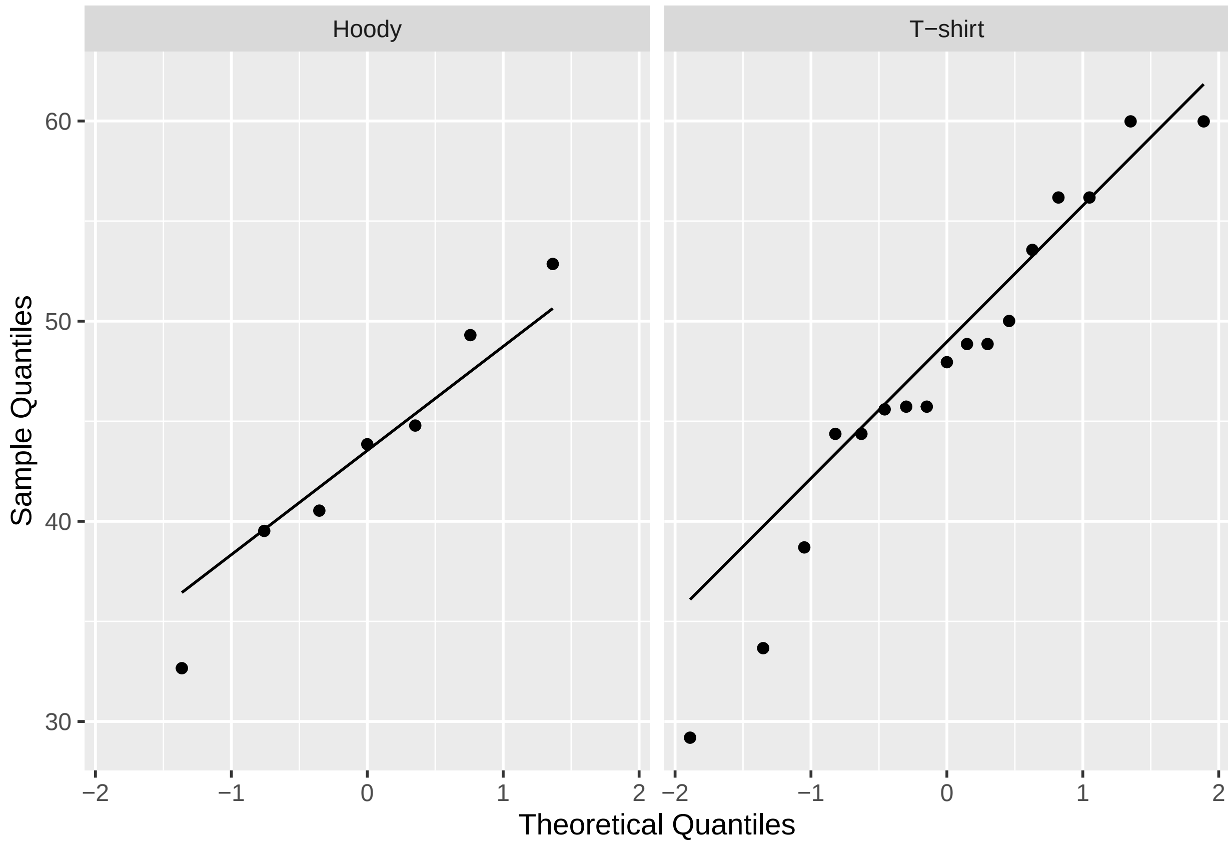
Example: Suppose we want test whether there is a significant difference in earnings per hour between Hoody products and T-shirt products of Outer Known customer group at 5% significance level.

```
# Loading relevant data
two.sample.data <- Hypothesis.data %>% filter(Product.name %in% c("Hoody", "T-shirt"))
```

Step 1: Check whether Earnings per hour values are normally distributed

Normal probability plot

```
ggplot(two.sample.data, aes(sample = Earnings.per.hour)) + stat_qq() +
  stat_qq_line() + facet_grid(.~Product.name) +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles")
```



Normality test

```
test1 <- two.sample.data %>% filter(Product.name == "Hoody")
shapiro.test(test1$Earnings.per.hour)
```

```

	Shapiro-Wilk normality test

data:  test1$Earnings.per.hour
W = 0.98413, p-value = 0.9771
```

```
test2 <- two.sample.data %>% filter(Product.name == "T-shirt")
shapiro.test(test2$Earnings.per.hour)
```

```

	Shapiro-Wilk normality test

data:  test2$Earnings.per.hour
W = 0.94884, p-value = 0.4384
```

Hypothesis to be tested:

H0: Data are normally distributed.

H1: Data are not normally distributed.

According to the Shapiro-Wilk normality test both p-values > 0.05.

Hence, We can conclude that Earnings per hour values of the two categories are normally distributed.

Step 2: Check for equality of variance

```
var.test(Earnings.per.hour ~ Product.name, data = two.sample.data,
         alternative = "two.sided")
```

```
    F test to compare two variances

data:  Earnings.per.hour by Product.name
F = 0.61833, num df = 6, denom df = 16, p-value = 0.5739
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1850935 3.2424317
sample estimates:
ratio of variances
         0.6183291
```

Hypothesis to be tested:

H0: Two population variances are equal.

H1: Two population variances are not equal.

According to the F test both p-values $= 0.5739 > 0.05$.

Hence, We can conclude that Two population variances are equal.

Step 3: Perform the t-test

```
t.test(Earnings.per.hour ~ Product.name, data = two.sample.data,
       alternative = "two.sided", var.equal = TRUE)
```

```
    Two Sample t-test

data:  Earnings.per.hour by Product.name
t = -1.1755, df = 22, p-value = 0.2524
alternative hypothesis: true difference in means between group Hoody and group T-shirt is not equal to 0
95 percent confidence interval:
 -11.672730   3.227215
sample estimates:
  mean in group Hoody mean in group T-shirt
             43.35860              47.58136
```

Since p-value $= 0.2524 > 0.05$, we do not reject null hypothesis.

Hence, there is sufficient evidence to conclude that there is a significant difference in earnings per hour between the two product types.

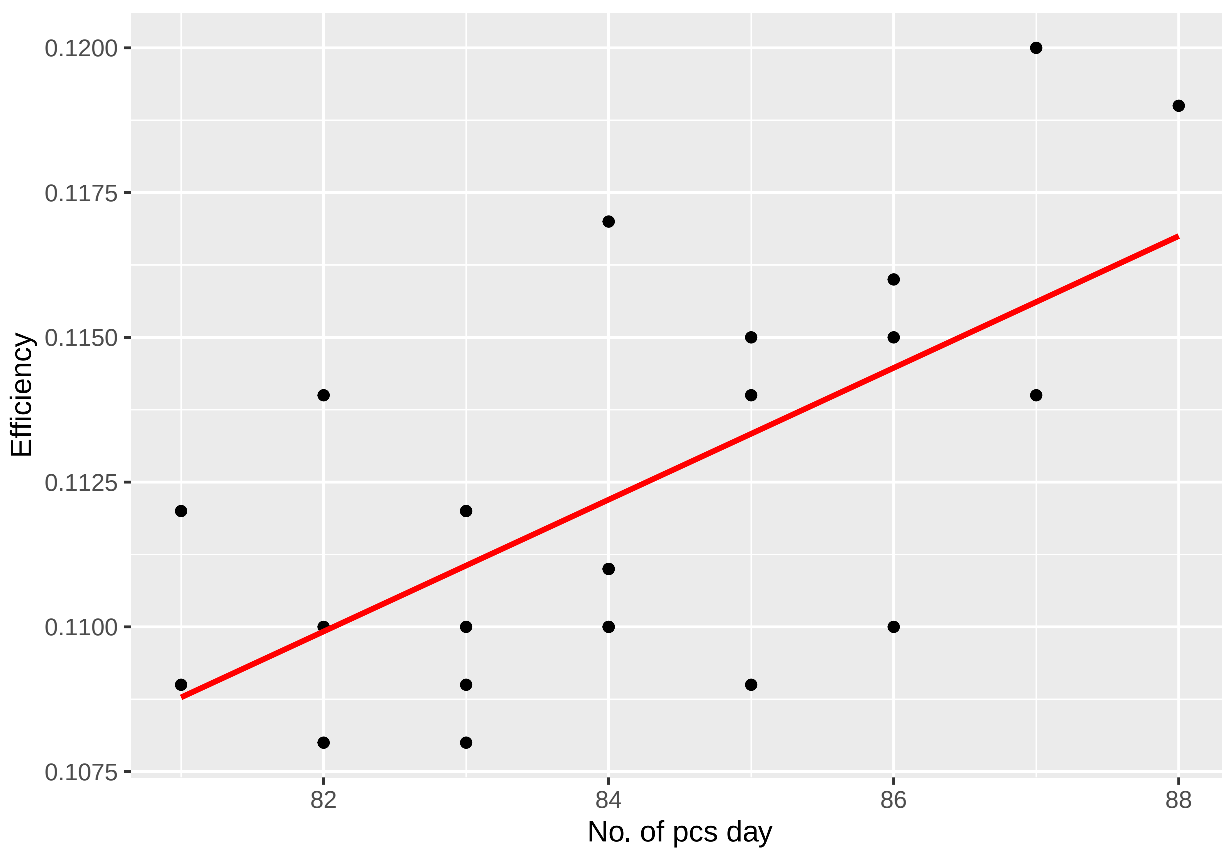
Simple linear regression analysis - Slide no 83

Example:
Suppose we aim to identify the factors affecting line efficiency in the apparel industry. To examine the relationship between selected variables and efficiency, we will conduct a multiple regression analysis. For this analysis, we will utilize the variables efficiency (Efficiency), and number of pieces per day (No. of pcs day). The response variable is efficiency whereas number of pieces per day is the predictor variables.

```
# load the data set
Reg_data <- read_excel("Textile.xlsx")
```

Check the linearity assumption

```
# scatter plot of Earnings and Standard Hours

ggplot(Reg_data, aes(x=`No. of pcs day`, y=Efficiency)) +
  geom_point() + geom_smooth(method = lm, se = FALSE, color = "red")
```




```
  ggtitle("Scatterplot of Efficiency and No.of pieces per day")
```

```
$title
[1] "Scatterplot of Efficiency and No.of pieces per day"
```

```
attr(,"class")
[1] "labels"
```

Fit the model

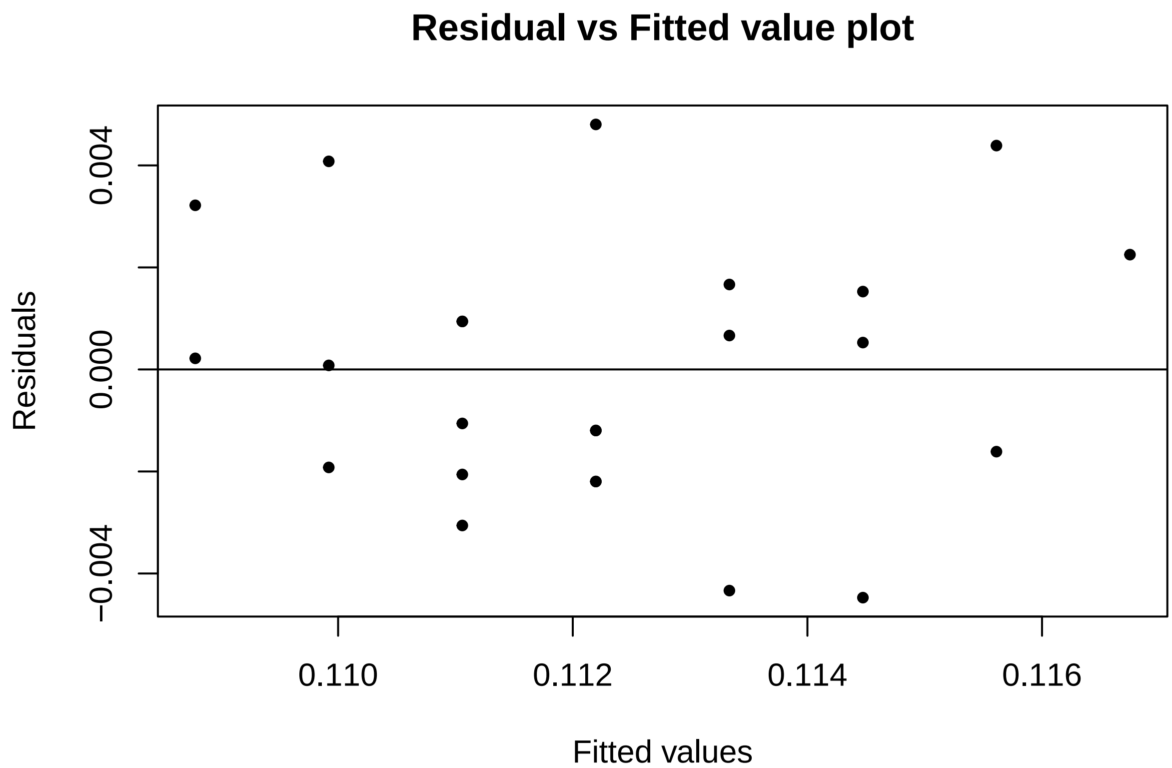
```
Call:
lm(formula = Efficiency ~ `No. of pcs day`, data = Reg_data)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0044732 -0.0019550  0.0001486  0.0015614  0.0048032 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      0.0165905  0.0243311   0.682 0.502442    
`No. of pcs day` 0.0011382  0.0002893   3.934 0.000708 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002649 on 22 degrees of freedom
Multiple R-squared:  0.413, Adjusted R-squared:  0.3863 
F-statistic: 15.48 on 1 and 22 DF,  p-value: 0.0007079
```

According to the R^2 value, it can be said that 38.6% of variation in efficiency can be explained by the fitted model.

Checking assumptions

Check the constant variance assumption of residuals

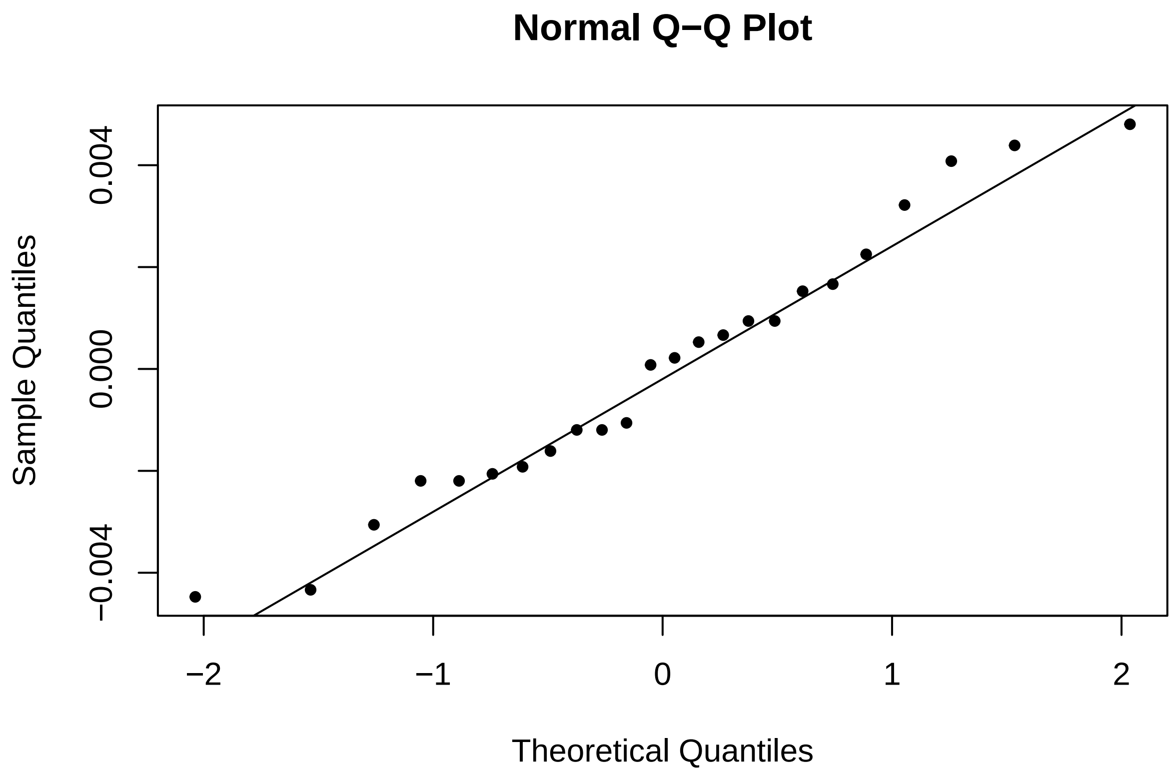
```r
# obtain the residuals
residuals <- resid(model1)

plot(fitted(model1), residuals, xlab="Fitted values",ylab="Residuals",
     main="Residual vs Fitted value plot", pch=20, cex=1)


# add a horizontal line at 0
abline(0,0)
```



According to the residual vs fitted value plot, it can be seen that data are scattered around 0 without any systematic pattern. So, it can be concluded that the residuals are independent and has a constant variance. Therefore, it can be concluded that model adequately fit the data.

Check the normal assumption of residuals

```
# Q-Q plot for residuals
qqnorm(residuals, pch=20)

# add a straight diagonal line to the plot
qqline(residuals)
```




```
# Normality test
shapiro.test(residuals)
```

```

	Shapiro-Wilk normality test

data:  residuals
W = 0.96876, p-value = 0.6365
```

Hypothesis to be tested:

H0: Residuals are normally distributed.

H1: Residuals are not normally distributed.

According to the Shapiro-Wilk normality test p-value $= 0.3724 > 0.05$.

Hence, We can conclude that residuals are normally distributed.

Final fitted model for efficiency

Efficiency $= 0.016 + 0.001$ No. of pieces day

The model shows that the number of pieces per day was positively related with efficiency. However, the contribution from this variable to model was relatively small. This may be due to some other factors which are not considered here, that affect efficiency.