VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

PROBABILITY AND STATISTIC

Report

Advisors: Nguyễn Tiến Dũng

Class code: CC09

Students: Nguyễn Đức Thành - 1952983
Cao Bá Huy - 1952713
Đỗ Đăng Khoa - 1952295
Phạm Quang Khánh - 1852459

HO CHI MINH CITY, MAY 2021

Contents

1 Project 1

1.1 Problem 1

1.1.1 Requirement

In order to compare advertising costs in four different newspapers (with the same advertising conditions), a sample of 7 advertising articles was collected from each newspaper and the following results (in thousand VND).

Newspaper A	57	65	50	45	70	62	48
Newspaper B	72	81	64	55	90	38	75
Newspaper C	35	42	58	59	46	60	61
Newspaper D	73	85	92	68	82	94	66

Find the p-value to determine if there is any significant difference in advertising costs among these newspapers.

1.1.2 Method

Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

1.1.3 Implementation

Step 1 : Set up the hypothesis
- H_0 : The average advertising costs in four different newspapers is the same.
- H_1 : The average advertising costs in four different newspapers is different.

Step 2 : Using R code and result

```
# Input data
data = matrix(c(57, 65, 50, 45, 70, 62, 48, 72, 81, 64, 55, 90, 38, 75, 35,
↪  42, 58, 59, 46, 60, 61, 73, 85, 92, 68, 82, 94, 66), ncol = 7, byrow =
↪  TRUE)

# Set name for 4 kind of newpapers
rownames(data) = c("Newpapers A", "Newpapers B", "Newpapers C", "Newpapers
↪  D")
```

```r
# Calculate p-value
chisq.test(data)
```

Result:

```
        Pearson's Chi-squared test

data:  data
X-squared = 47.943, df = 18, p-value = 0.0001535
```

Step 3 : Compare with the significant difference
Due to p-value = 0.01525%, thus we reject assumption H_0

1.1.4 Conclusion

The advertising costs among four different newpapers is the same

1.2 Problem 2

1.2.1 Requirement

The following table gives the counts of hair color of a random sample of people.

Hair color	Male	Female
Black	56	32
Red	37	66
Brown	84	90
Yellow	19	38

At the significance level of 3%, determine whether hair color and gender are correlated.

1.2.2 Method

The Chi-Square Test is used to determine if there is any association between two variables.

1.2.3 Implementation

Step 1 : Set up the hypothesis
- H_0 : There is no association between hair color and gender.
- H_1 : Hair color and gender are correlated.

Step 2 : Using R code and result

```
# create data for male
Male <- c(56,37,84,19)
# create data for female
Female <- c(32,66,90,38)
# create the whole data
Data <- data.frame(Male,Female)
# test the data
chisq.test(Data)
```

Result:

```
        Pearson's Chi-squared test

data:  Data
X-squared = 19.215, df = 3, p-value = 0.0002468
```

Step 3 : Compare with the significant difference

In these results, the Pearson's Chi-square statistic is 19.215 and the p-value(≈ 0.00025) $<$ $\alpha(=0.03)$. Therefore, at a significance level of 3%, we reject assumption H_0 and fail to reject assumption H_1 that hair color and gender are associated.

1.2.4 Conclusion

There is an association between hair color and gender.

1.3 Problem 3

1.3.1 Requirement

In order to check whether the primary occupations and secondary occupations affect the average income of households, the following record of average incomes is collected.

Primary occupations	Secondary occupations			
	(1)	(2)	(3)	(4)
Rice cultivation	3.5	7.4	8.0	3.5
Fruit farming	5.6	4.1	6.1	9.6
Breeding	4.1	2.5	1.8	2.1
Services	7.2	3.2	2.2	1.5

Draw a conclusion at the significance level of 3%.

1.3.2 Method

In this problem, we need to determine if the two independent variabes , which are primary occupations and secondary occupations have the effect on the dependent variable which is the average income. Therefore, we will use the **2-way ANOVA** method to implement the checking.

1.3.3 Implementation

Step 1 : Set up the hypothesis
- H_{0A} : The primary occupations does not effect the average income of households.
- H_{1A} : The primary occupations affects the average income of households.
- H_{0B} : The secondary occupations does not effect the average income of households.
- H_{1A} : The secondary occupations affects the average income of households.

Step 2 : R code

```
#Generate the data
primary<- gl(4,4,16,labels = c("Rice cultivation" , "Fruit farming" ,
                               "Breeding" , "Services"))
secondary <- gl(4,1,16)
income <- c(3.5, 7.4, 8.0, 3.5, 5.6, 4.1, 6.1, 9.6,
            4.1, 2.5, 1.8, 2.1, 7.2, 3.2, 2.2, 1.5)
list <- data.frame(primary,secondary,income)
# Anova test for the effect of jobs on income
two.way <- aov(income ~ primary + secondary, data = list)
```

```
summary(two.way)
```

Then, we have the output of ANOVA test :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
primary	3	36.39	12.128	1.997	0.185
secondary	3	2.01	0.672	0.111	0.952
Residuals	9	54.67	6.074		

Step 3 : Find the F value and compare
With the significance level of 3 %, we have to calculate the value of quantile function over F distribution:

```
> qf(1-0.03,3,9)
[1] 4.740652
```

Moreover , the primary and secondary job have the same degree of freedom (3) , thereby they have the same F_{test} value which is 4.740652.

Apparently, we have :

$$F_{primary} = 1.9966 < F_{test} = 4.740652$$
$$F_{secondary} = 0.1106 < F_{test} = 4.740652$$

We can say that both value F for Primary and secondary occupations are outside the reject region. Therefore, we do not reject the null hypothesis.

1.3.4 Conclusion

By using the 2-way ANOVA method to analyse, we can conclude that the primary occupations and secondary occupations do not affect the average income of households .

1.4 Problem 4

1.4.1 Requirement

The saponin content (mg) of medicinal plant of the same type that were harvested at different times of seasons in three regions is given the following table.

Seasons	Time	Regions		
		South	Central	North
Dry season	Early	2.4	2.1	3.2
	Mid	2.4	2.2	3.2
	Late	2.5	2.2	3.4
Rainy	Early	2.5	2.2	3.4
	Mid	2.5	2.3	3.5
	Late	2.6	2.3	3.5

Is there any significant difference in the saponin content among the seasons and the regions? Do the two factors season and region interact? Use the significance level of 5

1.4.2 Method

In order to determine if there is a significant difference in the saponin content among the seasons and the regions as well as the interaction between the two factors, we need to perform a 2-way ANOVA test to compare the F values and draw the conclusions from there.

1.4.3 Implementation

Step 1: Set up the hypotheses
- H_{0A}: There is no significant difference in the saponin content among the seasons.
- H_{1A}: There is a significant difference in the saponin content among the seasons.
- H_{0B}: There is no significant difference in the saponin content among the regions.
- H_{1B}: There is a significant difference in the saponin content among the seasons.
- H_{0AB}: There is no interaction between the two factors season and region.
- H_{1AB}: There is a interaction between the two factors season and region.

Step 2: Use R to perform the ANOVA test

```
#Preparing the data
alpha = 0.05
Seasons <- gl(2, 9, 18, labels = c("Dry", "Rainy"))
Regions <- gl(3, 3, 18, labels = c("South", "Central", "North"))
```

```
Saponin_content <- c(2.4, 2.4, 2.5, 2.1, 2.2, 2.2, 3.2, 3.2, 3.4, 2.5,
2.5, 2.6, 2.2, 2.3, 2.3, 3.4, 3.5, 3.5)

#Performing 2-way ANOVA
data <- data.frame(Seasons, Regions, Saponin_content)
res <- aov(data = data, Saponin_content ~ Seasons * Regions)
print(summary(res))
```

The result of the 2-way ANOVA test is as follows:

```
                Df  Sum Sq  Mean Sq  F value  Pr(>F)
Seasons          1   0.080    0.080       16  0.00176     **
Regions          2   4.348    2.174    434.8  6.36e-12    ***
Seasons:Regions  2   0.010    0.005      1.0  0.39657
Residuals       12   0.060    0.005
```

Note that $F_A = 16$, $F_B = 434.8$ and $F_{AB} = 1.0$ for the respective F-values of the proposed factors: season, region and season * region.

Step 3: Calculate the F values and compare
Given the significance level of 5%, we can calculate the F-value associated with the specified cumulative probability distribution:

```
cat("F_a =", F_a <- qf(1-alpha,1,12), "\n")
cat("F_b =", F_b <- qf(1-alpha,2,12), "\n")
cat("F_ab =", F_ab <- qf(1-alpha,2,12), "\n")
```

Thus we obtain the F-values needed to compare with the F-values returned from the ANOVA test:

```
F_a = 4.747225
F_b = 3.885294
F_ab = 3.885294
```

Here, we compare the F-values from the ANOVA test with the F-values calculated from the significance level.

$$F_A = 16 > F_a = 4.75$$
$$F_B = 434.8 > F_b = 3.89$$
$$F_{AB} = 1 < F_{ab} = 3.89$$

1.4.4 Conclusion

From the comparisons above, the appropriate course of action is:
- To reject H_{0A}, meaning there is a significant difference in the saponin content among the

seasons.
- To reject H_{0B}, meaning there is a significant difference in the saponin content among the regions.
- To not reject H_{0AB}, meaning there is no interaction between the two factors season and region regarding the saponin content.

2 Project 2

2.1 Introduction

2.2 Data Interpretation

2.2.1 Data description

The data set called flights contains information about 162049 flights that departed from the two major airports of the Pacific Northwest (PNW), SEA in Seattle and PDX in Portland, in 2014. Each is represented by a row.
The attributes (columns) in our data include:

- *year, month, day* : date of the departure where year is 2014 by deafult
- *carrier* : the flight carrier
- *origin* : the departure airport with SEA in Seattle and PDX in Portland
- *dest* : the destination airport
- *dep_time* : estimated time departure
- *arr_time* : estimated arrival departure
- *dep_delay* : departure delay
- *arr_dalay* : arrival delay
- *distance* : distance between two airports (in miles)

2.2.2 Data cleaning

However the data set is missing some values for some attributes which are tagged as NA (Not available). So it is necessary to clean those NA values. In this project, we replace the NA value with its respective attribute mean.

```
1  #function to replace all NA with respective column mean
2  replace_NA <- function(fr)
3  {
4    new_frame <- fr
5    num_col <- ncol(new_frame)
6    for (x in 1:num_col)
7    {
8      if (is.numeric(new_frame[, x]))
9      {
10       avg <- mean(new_frame[, x], na.rm = TRUE)
11       new_frame[is.na(new_frame[, x]), x] <- avg
12     }
```

```
  }
  eval.parent(substitute(fr <- new_frame))
}
```

2.2.3 Statistical overview

For numeric attributes
We will do some basic statistic to get the mean, standard variation , min and max of each attributes in order to have an overview about our data set. Below is the function we use to get the basic information of our numeric attributes:

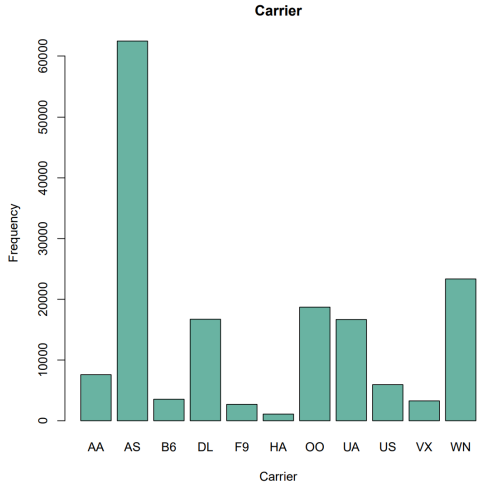
```
basic_statistic <- function(){
  cat("For numeric attributes \n")
  print("For departure time", quote = FALSE )
  print(summary(flights$dep_time))
  cat("Standard deviation : ", sd( flights$dep_time ),"\n \n")
  print("For departure delay", quote = FALSE )
  print(summary(flights$dep_delay))
  cat("Standard deviation : ", sd( flights$dep_delay ),"\n\n")
  print("For arrival time", quote = FALSE )
  print(summary(flights$arr_time))
  cat("Standard deviation : ", sd( flights$arr_time ),"\n\n")
  print("For arival delay", quote = FALSE )
  print(summary(flights$arr_delay))
  cat("Standard deviation : ", sd( flights$arr_delay ),"\n\n")
  print("For flight distance", quote = FALSE )
  print(summary(flights$distance))
  cat("Standard deviation : ", sd( flights$distance ),"\n")
  cat("For carrier \n")
  print(as.data.frame(table(flights$carrier)))
  cat("\nFor origin \n")
  print(as.data.frame(table(flights$origin)))
  cat("\n For Destination \n ")
  print(as.data.frame(table(flights$dest)))
}
```

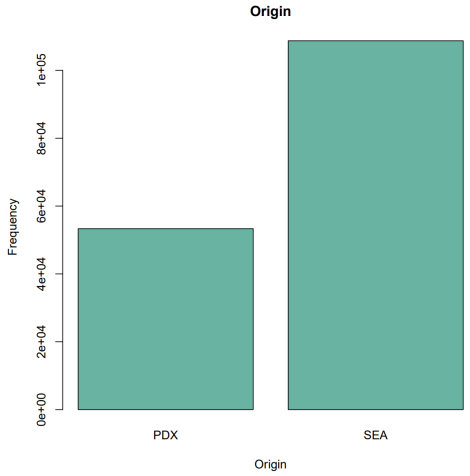
Then we will summarize the output with the below table :

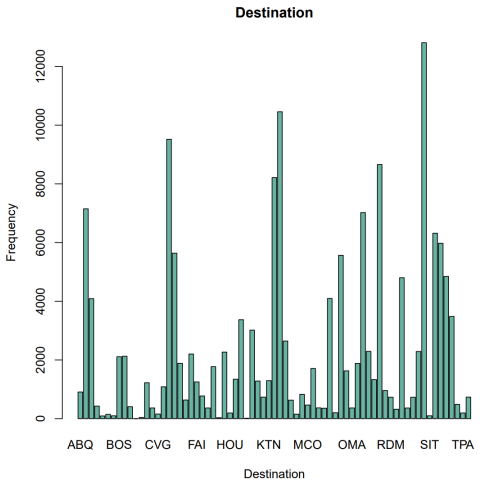
	dep_time	dep_delay	arr_time	arr_delay	distance
Mean	1278	6.134	1483	2.241	1205
SD	521.2	29.035	522.359	31.07	653.15
Min	1	-37	1	-67	93
Max	2400	1553	2400	1539	2724

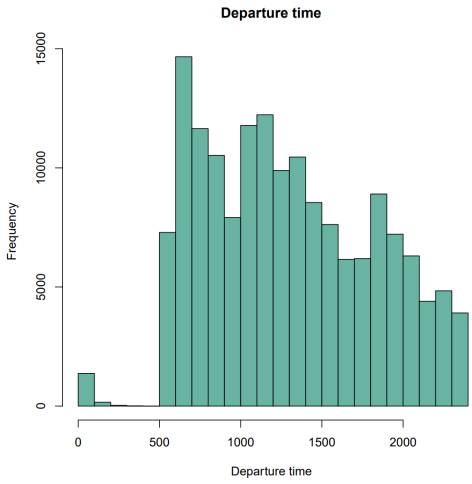
2.2.4 Data visualization

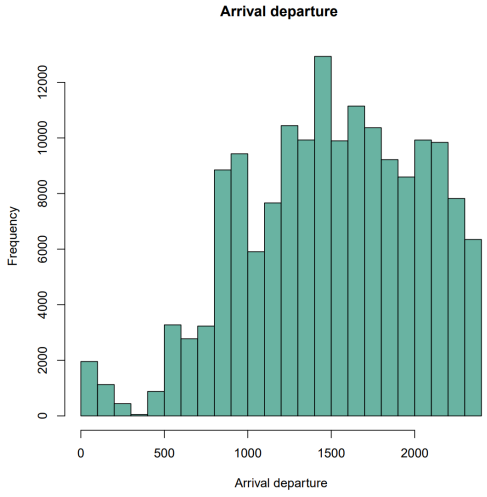
We will have a distribution visual for each attribute.

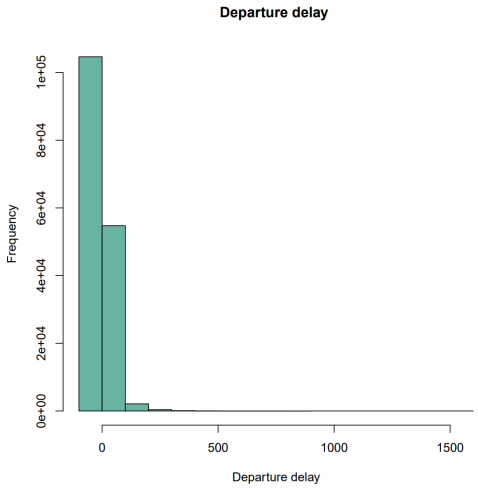


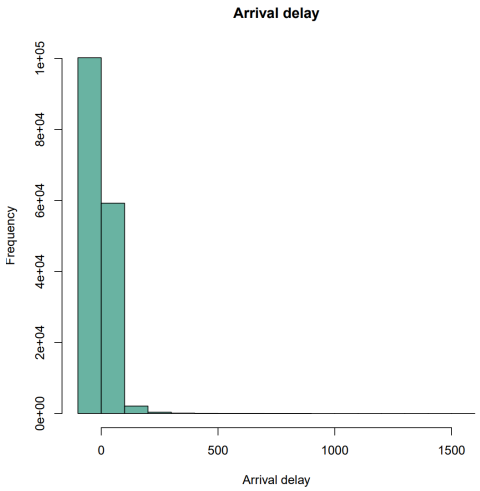


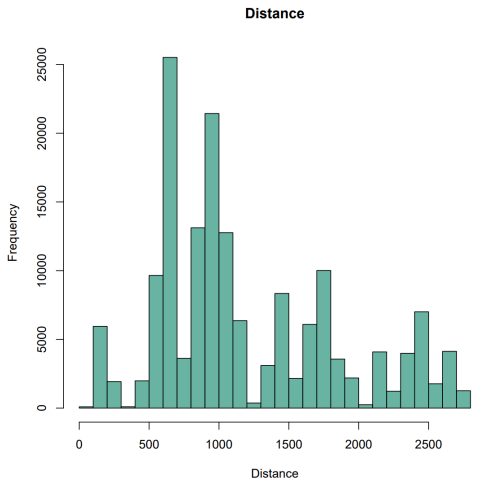





Arrival departure
Frequency
0
2000
4000
6000
8000
10000
12000
0
500
1000
1500
2000
Arrival departure


Departure delay
Frequency
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
0
500
1000
1500
Departure delay


Arrival delay
Frequency
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
0
500
1000
1500
Arrival delay


Distance
Frequency
0
5000
10000
15000
20000
25000
0
500
1000
1500
2000
2500
Distance

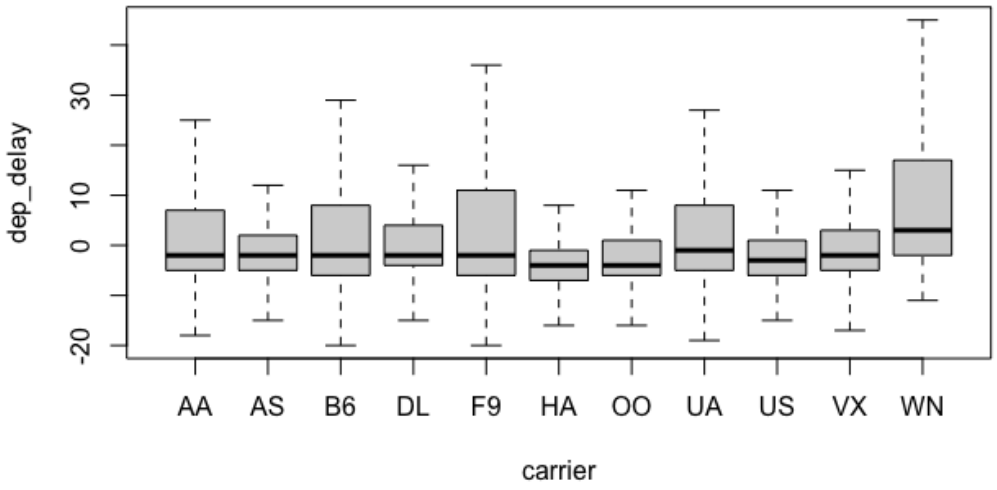
2.2.5 Graphs: boxplot - dep_delay for each carrier.

First, we split the data of dep_delay by the carrier.
Then, we will use the function boxplot in R to plot the data of dep_delay for each carrier, set outline = FALSE to remove all outliers.

R code:

```
ploted_data <- split(flights$dep_delay,flights$carrier)
boxplot(ploted_data,xlab="carrier",ylab="dep_delay", outline=FALSE)
```

Here is the result:



2.3 Oneway ANOVA

2.3.1 Set up the hypothesis

- H_0: The average departure delays are the same for all airlines for flights departing from Portland in 2014.
- H_1: At least one of the mean departure delays is different.

2.3.2 Get the data

From the given data, we will choose all the flights in 2014 that are departed from Portland.
R code:

```
data <- subset(flights, year == 2014 & origin == "PDX")
```

2.3.3 Explore the data

R code:

```
library(knitr)
data_summ <- data \%>\% group_by(carrier) \%>\%
summarize(sample_size = n(), mean = mean(dep_delay), sd = sd(dep_delay),
↪  minimum = min(dep_delay), max = max(dep_delay))
kable(data_summ)
```

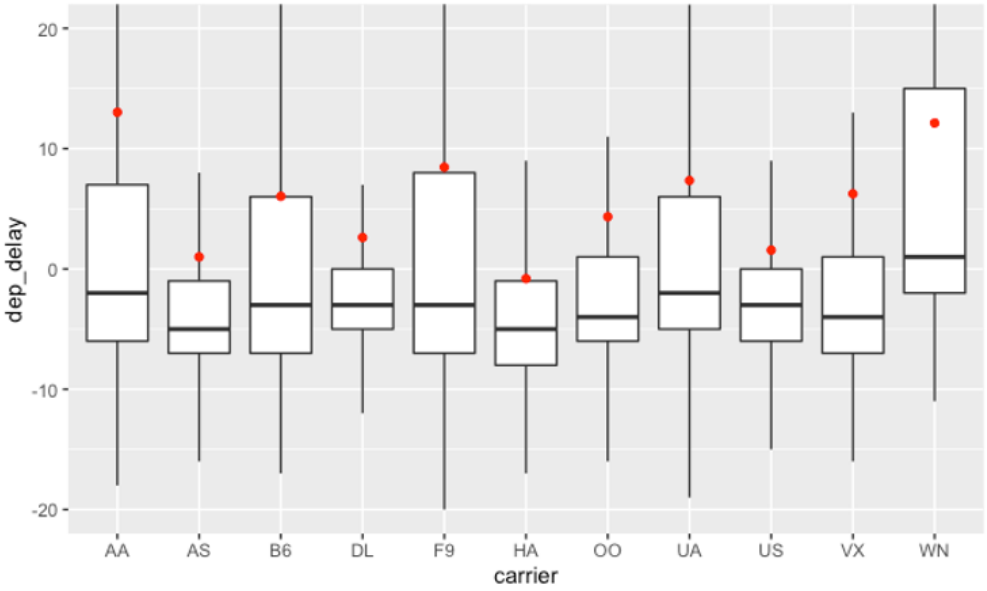
Result:

carrier	sample_size	mean	sd	minimum	max
AA	2187	13.0277157	66.56780	-18	1553
AS	12844	1.0003176	22.72195	-25	340
B6	1287	6.0362131	29.48320	-17	265
DL	5168	2.6172616	28.07475	-19	648
F9	1362	8.4608438	38.08966	-20	590
HA	365	-0.8027397	25.28007	-17	417
OO	9841	4.3383699	27.25609	-18	411
UA	6061	7.3520806	30.89081	-19	486
US	2361	1.5646813	23.36553	-17	346
VX	666	6.2477477	33.74758	-21	358
WN	11193	12.1305118	29.96098	-11	712

Plot the boxplot and show the mean of the *dep_delay* for each carrier highlighted by the red dots:

R code:

```
library(ggplot2)
ggplot(aes(x = carrier, y = dep_delay), data = data) +
↪  geom_boxplot(outlier.shape = NA) + coord_cartesian(ylim = c(-20, 20)) +
↪  stat_summary(fun = "mean", geom = "point", color = "red")
```



2.3.4 Compute the p_value:

R code to calculate the One-way ANOVA test:

```
res.aov <- aov(dep_delay ~ carrier, data = data)
summary(res.aov)
```

Result

```
                Df   Sum Sq Mean Sq F value Pr(>F)
carrier         10  1014392  101439   110.6 <2e-16 ***
Residuals   53324 48900289     917
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2.3.5 Draw conclusion

With the very tiny p-value (0), therefore, we have sufficient evidence to reject the null hypothesis. At significant level of almost 0%, we could state that at least one of the population mean departure delays is different.

2.4 Linear model

2.4.1 Requirement

In this section, we will have to generalize a linear model to evaluate how *dep_delay* and carrier affect the *arr_delay*

2.4.2 Method

Regression analysis is a collection of statistical tools that are used to model and explore relationships between variables that are related in a non-deterministic manner.
Multiple Linear Regression (MLR) attempts to model the linear relationship between a dependent variable (response) and some independent variables (predictors/regressors). A model that might describe this relationship is :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Where $\beta_0, \beta_1, ..., \beta_n$ are called partial regression coefficients since β_i measures the change of Y per unit change in x_i when other variables remain constant.
In this project, *arr_delay* is the dependent variable , while *dep_delay* and carrier are the independent variables.

However, regression analysis requires numerical variables while the carrier is a categorical variable (also known as factor or qualitative variables) that has 11 levels representing the 11 carrier types in the flights dataset. Therefore, if we want to include the categorical variable in our model, the categorical variable has to be recoded into a set of separate dichotomous variables. This recoding is called **"dummy coding"** and leads to the creation of a table called **contrast matrix**.

Generally, we will transform 11 levels from carier into 10 variables taking a 0 or 1. These 10 new variables contain the same information as the single variable.
We will generate a logical matrix with the column name representing the name of the each flight carriers : $AS, US, UA, DL, AA, F9, VX, OO, WN, B6$. For example :

	dep_delay	arr_delay	AS	US	UA	DL	AA	F9	VX	OO	WN	B6
1	96	70	1	0	0	0	0	0	0	0	0	0

The number 1 indicates the carrier of the flight, so in the above example, the flight carrier is US. However, there is another carrier called HA that has not been included in the matrix yet , so in order to indicate the flight carrier is HA, all the dummies variable are set to 0. For example:

39	12	0	0	0	0	0	0	0	0	0	0	0

When setting the matrix is done, we are ready to generalize our linear model.

2.4.3 Implementation R code

First of all, we need to prepare the data frame and the logical matrix.

```
carrier_list <- unique(flights$carrier)
num_carrier <- length(carrier_list) # the number of carriers
dvar_colname <- c()
#create a vector from x1 to x10
for (x in 1:(num_carrier - 1))
{
  dvar_colname <- append(dvar_colname, carrier_list[x])
}
print(dvar_colname)
# create a separate frame
dep_delay <- flights$dep_delay
arr_delay <- flights$arr_delay
lm_frame <- data.frame(dep_delay, arr_delay)
#convert carrier factor to dummy vars
for (x in 1:(num_carrier - 1))
{
  new_col <- ifelse(flights$carrier == carrier_list[x], 1, 0)
  lm_frame[dvar_colname[x]] <- new_col
}
```

After this, we will generalize the linear model with *arr_delay* as the dependent variable and *dep_delay* and carrier as the independent variables:

```
res <- lm(arr_delay ~ dep_delay + AS + US + UA + DL + AA + F9 +VX + OO + WN + B6, data = lm_frame)
#shouldn't it be this, i mean syntactically the latter makes more sense and is more in line with what we already have
#R studio probably provides the context for the vars but eh, the latter is universally correct
res <- lm(lm_frame$arr_delay ~ lm_frame$dep_delay + lm_frame$x1 + lm_frame$x2 + lm_frame$x3 + lm_frame$x4 + lm_frame$x5
```

2.4.4 Result

Finally we will evaluate the result :

```
1 print(summary(res))
```

Output :

```
Residuals:
    Min      1Q  Median      3Q     Max
-321.43   -6.93   -0.39    6.33  167.36

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.498601   0.363375  -1.372  0.17002
dep_delay    0.990300   0.001038 953.777  < 2e-16 ***
AS          -2.163377   0.366537  -5.902 3.59e-09 ***
US          -3.525945   0.395411  -8.917  < 2e-16 ***
UA          -6.788058   0.375183 -18.093  < 2e-16 ***
DL          -4.607029   0.375085 -12.283  < 2e-16 ***
AA          -4.227899   0.388795 -10.874  < 2e-16 ***
F9          -0.281325   0.430909  -0.653  0.51385
VX          -3.583333   0.419822  -8.535  < 2e-16 ***
OO          -1.143223   0.373852  -3.058  0.00223 **
WN          -4.873070   0.371953 -13.101  < 2e-16 ***
B6          -4.066182   0.415827  -9.779  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.02 on 162037 degrees of freedom
Multiple R-squared:  0.8502,    Adjusted R-squared:  0.8502
F-statistic: 8.36e+04 on 11 and 162037 DF,  p-value: < 2.2e-16
```

From the result, we can conclude that -0.4986 is the expected mean value of arr$_d$*elaywhenalltheindependentvariablesa*
valueofcarrierF9is > 0.05*sothecarrierF9doesnotaffectthe*arr_delay*.Ontheotherhand, othercarriersand*dep_dela
*Next, wedescribetherelationshipof*arr_delay, dep_delay*andeachofcarrierfromtheaboveresultasfollows* :

For carrier AS : arr_delay = -0.4986 + 0.99*dep_delay - 2.163 .

For carrier US : arr_delay = -0.4986 + 0.99*dep_delay - 3.526 .

For carrier UA : arr_delay = -0.4986 + 0.99*dep_delay - 6.788 .

For carrier DL : arr_delay = -0.4986 + 0.99*dep_delay - 4.607 .

For carrier AA : arr_delay = -0.4986 + 0.99*dep_delay - 4.227 .

For carrier VX : arr_delay = -0.4986 + 0.99*dep_delay - 3.583 .

For carrier OO : arr_delay = -0.4986 + 0.99*dep_delay - 1.143 .

For carrier WN : arr_delay = -0.4986 + 0.99*dep_delay - 4.873 .

For carrier B6 : arr_delay = -0.4986 + 0.99*dep_delay - 4.066 .

For carrier HA : arr_delay = -0.4986 + 0.99*dep_delay .