Name – Rahul Sharma

Basic Statistics tasks

1. Which store has maximum sales

```
getwd()

library(readr)

Walmart_Store_sales
<-
read_csv("C:/Users/Ra
hul/Downloads/Walma
rt_Store_sales.csv")

df <-
Walmart_Store_sales
view(df)
summary(df)

colnames(df)

attach(df)

library(dplyr)

library(magrittr)

summary_1 <-

df%>%

  group_by(Store)%>%


summarise(mean = mean(Weekly_Sales),

sum = sum(Weekly_Sales),

          max = max(Weekly_Sales),

          std = sd(Weekly_Sales))

summary_1

#Arranging the data

Arrange <- arrange(summary_1, desc(max))

Arrange

head(Arrange)
```

Output

```
# A tibble: 6 x 5
  Store     mean        sum      max     std
  <int>    <dbl>      <dbl>    <dbl>   <dbl>
1    14 2020978. 288999911. 3818686. 317570.
2    20 2107677. 301397792. 3766687. 275901.
3    10 1899425. 271617714. 3749058. 302262.
4     4 2094713. 299543953. 3676389. 266201.
5    13 2003620. 286517704. 3595903. 265507.
6     2 1925751. 275382441. 3436008. 237684.
```

Inference

Store 14 has the maximum sales.

2. Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation?

```
Arrange1 <- arrange(summary_1, desc(std))

Arrange1

head(Arrange1)
```

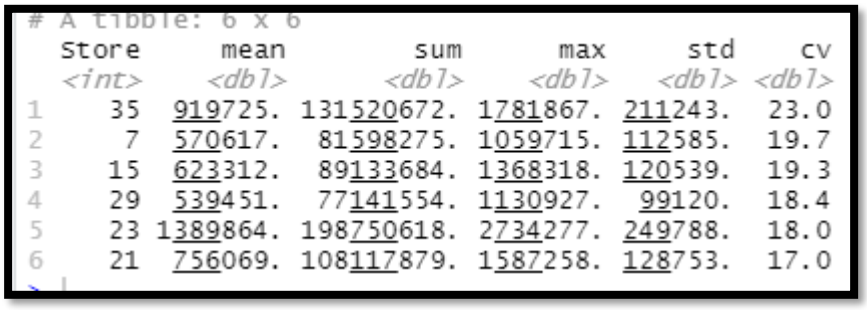
Output

```
# A tibble: 6 x 5
  Store     mean        sum      max     std
  <int>    <dbl>      <dbl>    <dbl>   <dbl>
1    14 2020978. 288999911. 3818686. 317570.
2    10 1899425. 271617714. 3749058. 302262.
3    20 2107677. 301397792. 3766687. 275901.
4     4 2094713. 299543953. 3676389. 266201.
5    13 2003620. 286517704. 3595903. 265507.
6    23 1389864. 198750618. 2734277. 249788.
```

```
summary_1$cv <- round(((summary_1$std/summary_1$mean)*100),2)

Arrange2 <- arrange(summary_1, desc(cv))

Arrange2
```

head(Arrange2)

Output



```
# A tibble: 6 x 6
  Store     mean        sum      max     std    cv
  <int>    <dbl>      <dbl>    <dbl>   <dbl> <dbl>
1    35  919725. 131520672. 1781867. 211243.  23.0
2     7  570617.  81598275. 1059715. 112585.  19.7
3    15  623312.  89133684. 1368318. 120539.  19.3
4    29  539451.  77141554. 1130927.  99120.  18.4
5    23 1389864. 198750618. 2734277. 249788.  18.0
6    21  756069. 108117879. 1587258. 128753.  17.0
```

Inference

Store 14 has the highest STD

Store 35 has the highest COV

3. Which store/s has good quarterly growth rate in Q3'2012

check data formats

sapply(df, class)

dtype <- sapply(df, class)

change the date format

df$date_v2 <- as.Date(df$Date, format = "%d-%m-%Y")

Check data format again

dtype <- sapply(df, class)

Take year and month

df$month_1 <- substring(df$Date,4,5)

df$year_1 <- substring(df$Date,7,10)

sapply(df, class)

change the format of month to numeric so I could use arithmatic operators for conditional statements

df$month_3 <- as.numeric(df$month_1)

attach(df)

```
# create a new variable called quarter
df$quarter[month_3 < 4] <- "Q1"
df$quarter[month_3 > 3 & month_3 <= 6] <- "Q2"
df$quarter[month_3 > 6 & month_3 <= 9] <- "Q3"
df$quarter[month_3 > 9] <- "Q4"
df$year_quarter <- transform(df, cat=interaction(year_1, quarter,sep='-'))
df$year_quarter <- paste(df$year_1,df$quarter, sep = "-")
# to get the sum of sales by store, year, and quarter
df2 <- aggregate(df$Weekly_Sales, by=list( store=df$Store, y_q=df$year_quarter), FUN=sum)
df2$store_y_q <- paste(df2$store,df2$y_q)
df2 <- arrange(df2, (store_y_q) )
attach(df2)
df3 <- select(df2, store, y_q, x)
install.packages("tidyverse")
library(tidyverse)
df3 <-
  df3 %>%
  rename(quarter_sale = x)
# Calculating my growth rate for stores
install.packages("plyr")
library(plyr)
df4 <- ddply(df3,"store",transform, growth=c(NA,exp(diff(log(x)))-1))
# rounding
df4$growth = round((df4$growth*100),1)
df5 <- filter(df4, y_q == "2012 Q3")
df5 <- arrange(df5, desc(growth) )
View(df5)
```

Output

store	y_q	x	growth
7	2012 Q3	8262787	13.3
16	2012 Q3	7121542	8.5
35	2012 Q3	11322421	4.5
26	2012 Q3	13675692	4.0
39	2012 Q3	20715116	2.5
41	2012 Q3	18093844	2.5
44	2012 Q3	4411251	2.4
24	2012 Q3	17976378	1.7
40	2012 Q3	12873195	1.1
23	2012 Q3	18641489	0.8
32	2012 Q3	15396529	-0.6
38	2012 Q3	5605482	-0.6

Inference

Store 7 has the highest quarterly growth rate in Q3 2012

4. Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together?

```
df_hd <- filter(df, Holiday_Flag == "1")
df_non_hd <- filter(df, Holiday_Flag != "1")
mean(df_non_hd$Weekly_Sales)
df_hd$sales <- as.numeric(df_hd$Weekly_Sales)
attach(df)
df_hd$flag[df_hd$sales > 1041256] <- "1"
df_hd2 <- filter(df_hd, flag == "1")
distinct(df_hd2, Date)
```

Output

	Date
1	12-02-2010
2	10-09-2010
3	26-11-2010
4	31-12-2010
5	11-02-2011
6	09-09-2011
7	25-11-2011
8	30-12-2011
9	10-02-2012
10	07-09-2012

Inference

Weekly sales for at least one store was greater than the average sales on non-holiday weeks for all stores for the below holidays:

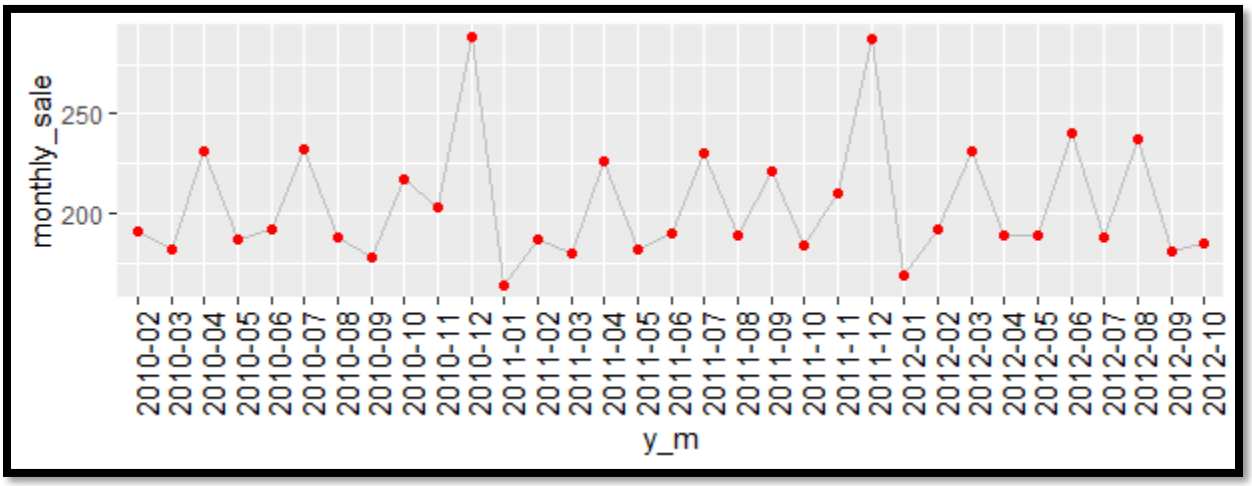
- Super Bowl week: 2010-2012
- Labour Day week: 2010-2012
- Thanksgiving week: 2010-2011
- Christmas week: 2010-2011

5. Provide a monthly and semester view of sales in units and give insights

 # monthly data

```
df$year_month <- format(as.Date(df$date_v2, format="%d-%m-%Y"),"%Y-%m")

df_m1 <- aggregate(df$Weekly_Sales, by=list( y_m=df$year_month), FUN=sum)

df_m1 <- arrange(df_m1, (y_m) )

df_m1$monthly_sale <- round((df_m1$x/1000000),2)

#df_m1$year_month_v2 <- as.Date(df_m1$y_m, format = "%Y-%m")

attach(df_m1)

plot(monthly_sale, type="l")

library(ggplot2)

ggplot(df_m1, aes(x=y_m, y=monthly_sale , group=1)) + geom_line(col="gray", linetype = "solid")+geom_point(col="red") +theme(axis.text.x=element_text(color = "black", size=11, angle=90, vjust=.8, hjust=0.8))
```

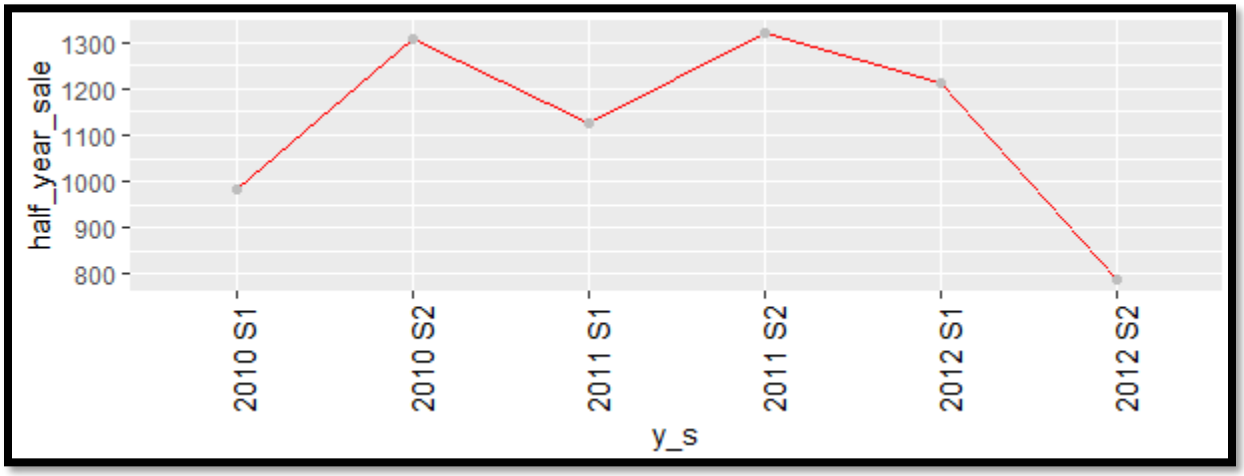
Monthly Data Output



 # Semester (half year)

```
df$semester[month_3 <= 6 ] <- "S1"
df$semester[month_3 > 6] <- "S2"
attach(df)
table(semester)
# semster data
df$year_semester <- paste(df$year_1,df$semester)
#Aggregate sales by semester
df_s1 <- aggregate(df$Weekly_Sales, by=list( y_s=df$year_semester), FUN=sum)
df_s1 <- arrange(df_s1, (y_s) )
df_s1$half_year_sale <- round((df_s1$x/1000000),2)
ggplot(df_s1, aes(x=y_s, y=half_year_sale , group=1)) + geom_line(col="red", linetype = "solid") + geom_point(col="gray") + theme(axis.text.x=element_text(color = "black", size=11, angle=90, vjust=.8, hjust=0.8))
```

Semester Data Output



Inference

There is increase in sales in the month of December, in line with question 4 the increase may due to Christmas holidays.

Sales also increase in the month of April & June.

Sales increased in the second half of the year in comparison with first half of the year.

Drastic decrease in the sales of 2nd semester could be because of incomplete data, so interpretation for second half of the year is not complete.

Statistical Model

For Store 1 – Build prediction models to forecast demand

```
df_store_1 <- filter(df, Store == "1")

df_store_1$date <- as.Date(df_store_1$Date, format = "%d-%m-%Y")

df_store_1 <- arrange(df_m1,(date)

df_store_1 <- cbind(date_new = rownames(df_store_1), df_store_1)

rownames(df_store_1) <- 1:nrow(df_store_1)

df_store_1$sales <- as.numeric(df_store_1$Weekly_Sales)

df_store_1$date_new <- as.numeric(df_store_1$date_new)

model_obj = lm(sales ~ date_new + CPI + Unemployment + Fuel_Price , data=df_store_1)

summary(model_obj)

AIC(model_obj)
```

Output – 1

```
Residuals:
    Min      1Q  Median      3Q     Max
-287031  -85237  -22986   61308  878829

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3483483.4  3001974.9  -1.160   0.2479
date_new          235.9     1426.9   0.165   0.8690
CPI             19855.8    13547.0   1.466   0.1450
Unemployment   124852.4    59178.7   2.110   0.0367 *
Fuel_Price     -67463.6    49616.7  -1.360   0.1761
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 151300 on 138 degrees of freedom
Multiple R-squared:  0.08517,   Adjusted R-squared:  0.05865
F-statistic: 3.212 on 4 and 138 DF,  p-value: 0.01479

> AIC(model_obj)
[1] 3823.926
```

model_obj2 = lm(sales ~ CPI + Unemployment + Fuel_Price + month_2 + year_1, data=df_store_1)

summary(model_obj2)

AIC(model_obj2)

Output – 2

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6130853    2203851  -2.782 0.006167 **
CPI             33167       9693   3.422 0.000821 ***
Unemployment    89571      79908   1.121 0.264283
Fuel_Price     -20383      63005  -0.324 0.746804
year_12011     -95648      65412  -1.462 0.145965
year_12012    -196575     120405  -1.633 0.104846
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 150400 on 137 degrees of freedom
Multiple R-squared:  0.1032,    Adjusted R-squared:  0.0705
F-statistic: 3.154 on 5 and 137 DF,  p-value: 0.009992

> AIC(model_obj2)
[1] 3823.074
```

Inference

The model was statistically significant with an R2 of 0.09

Only CPI was statistically significant

Other predictors are not statistically significant

Change dates into days by creating new variable

df$day <- format(as.Date(df$date_v2, format="%d-%m-%Y"),"%d")