Final Project

2023-03-08

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# load csv to R then save it as an object dataframe

# saveRDS(Cases, file = "Cases.RDS")
# saveRDS(Children, file = "Children.RDS")
# saveRDS(Payments, file = "Payments.RDS")
# saveRDS(Parents, file = "Parents.RDS")
```

```
# load dataframe
Cases <- readRDS("Cases.RDS")
Children <- readRDS("Children.RDS")
Payments <- readRDS("Payments.RDS")
Parents <- readRDS("Parents.RDS")
```

Question 1

a) Read the four CSV files into R, building four data frames with the names "Cases", "Parents", "Children" and "Payments". Show the dimensions of these data frames

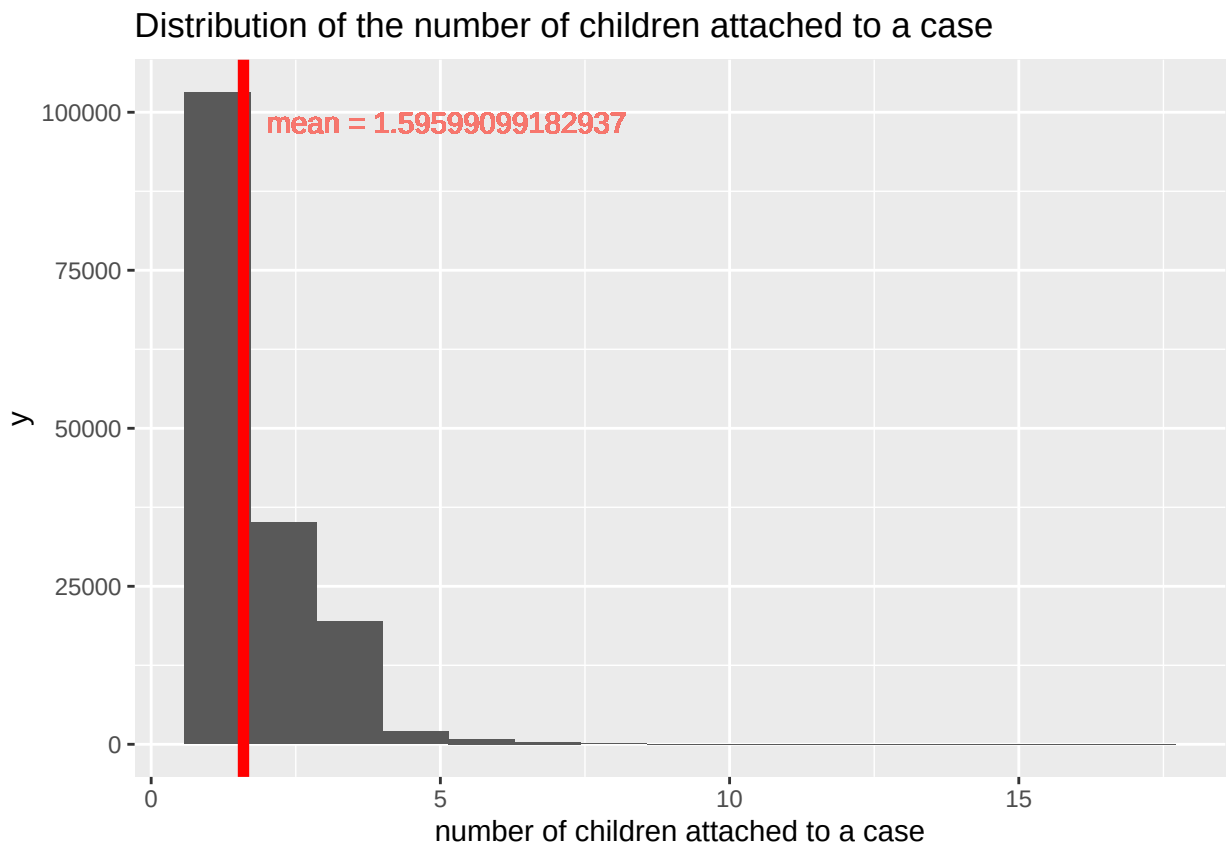
```r
# output the dimensions of 4 dataframes as a list

list("Dimension for Cases dataframe" = dim(Cases),
     "Dimension for Children dataframe" = dim(Children),
     "Dimension for Payments dataframe" = dim(Payments),
     "Dimension for Parents dataframe" = dim(Parents))
```

```
## $`Dimension for Cases dataframe`
## [1] 172422      6
##
## $`Dimension for Children dataframe`
## [1] 257253      9
##
## $`Dimension for Payments dataframe`
## [1] 1510216       6
##
## $`Dimension for Parents dataframe`
## [1] 128317     10
```

b) What is the distribution of the number of children attached to a case? Show an appropriate plot of the distribution, and mark the location of the average number in the plot.

The distribution is skewed right for the number of children attached to a case. The average number is approximately 2.

```r
# Count the cases of children that attached to a case
Children_Dist <- Children %>% group_by(CASE_NUM) %>% count(CASE_NUM)

# plot the counts and find the average number
ggplot(Children_Dist, aes(x = n))+
  geom_histogram(bins = 15)+
  geom_vline(xintercept = mean(Children_Dist$n), col = "red",size = 2) +
  geom_text(aes(x = 2, y = 100000, label = paste("mean =",mean(Children_Dist$n)),
            vjust = "inward",hjust = "inward", col = "#D22B2B")) +
  labs(
    title = "Distribution of the number of children attached to a case",
    x = "number of children attached to a case") +
  guides(color = "none")
```



c) The file children.csv may have more than one record for each child. What is the largest number of cases associated with a child, and indicate why you believe that this is indeed the same child.

Below is the largest number of cases associated with a child and the unique ID of that child. The output below show us that 12 is the largest number of cases associated with a child, and there are 2 children whose each appear 12 different times on 24 different cases. The reason I believe this is the same children because each child has their own unique ID and the output below show that there only 1 specific ID associate with each 12 different cases out of 24 cases in total.

```
# find the highest number cases that associate with a child
Children %>% group_by(ID) %>% summarise(cases_associated_with_ID = n()) %>%
  slice_max(cases_associated_with_ID)
```

```
## # A tibble: 2 x 2
##          ID cases_associated_with_ID
##       <int>                    <int>
## 1 153343287                       12
## 2 215334590                       12
```

d) Does every absent parent (AP_ID) identified in the payments data have an identifying record in the parents data file?

After examining both data frame by merging both data frames by the unique AP_ID, I figured that the count of all the AP_ID in the payments data have an identifying record in the parents data file because

when I joined the payments data file to the parents data parents and count all the cases, the total of AP_ID in the joined data frame matched with the AP_ID in the payments data file originally. As shown below, there are 28579 AP_ID for both the merged data frame and the payments data frame itself.

```
# joined parents to payments by AP ID
Subset_Parents_Payments <- Payments %>% select(AP_ID) %>% left_join(Parents, by = "AP_ID")

# count the total record in payments by AP ID after joining parents data frame
count_record_in_payments <- Subset_Parents_Payments %>% group_by(AP_ID) %>%
  summarise(total_cases = n());
dim(count_record_in_payments)
```

```
## [1] 28579     2
```

```
# count the total AP ID in payments record alone
count_cases_of_parents <- Payments %>% group_by(AP_ID) %>%
  summarise(total_cases_in_parents = n());
dim(count_cases_of_parents)
```

```
## [1] 28579     2
```

Question 2

Write a function named "pool_categories" that re-codes a categorical variable into a "simpler" factor with fewer categories by pooling categories with counts below a threshold into a category labeled 'Other' (a factor level which your function should check does not already exist!). You might find the R function %in% useful for this exercise.

```
pool_categories <- function(df_col, threshold) {
  if(typeof(df_col) == "character"){
    if(!("Other" %in% df_col)){
  # getting the frequency table of the input
    freq_table <- table(df_col)
  # take out the categories that's below the threshold
    below_threshold <- freq_table[freq_table < threshold]
  # if the categories that's below the threshold is in the
  #  input column, then pool the category labeled 'Others'
    df_col <- ifelse(df_col %in% names(below_threshold),"Others", as.character(df_col))
    df_col <- factor(table(df_col))
    } else {
      stop("Categories name Other is already existed in this given input")
    }
    return(df_col)
  } else {
    stop("input has to be a character type")
  }
}

pool_categories(Payments$PYMNT_SRC, 10)
```

```
##       A       C       F       G       I       L  Others       S       U       W
##   69144    2092    6690     513   19762     120       9    4305   50574 1356858
```

```
##       X       Z
##      70      79
## Levels: 9 70 79 120 513 2092 4305 6690 19762 50574 69144 1356858
```

Question 3

a) Make a variable Payments$DATE which is a viable R date by converting the COLLECTION_DT variable. Use this variable to find :

(i) the range of dates of all payments

(ii) the percentage of the total number of payments made before May 1, 2015

```
#arrange all payments by collection date
payments2<- Payments%>% arrange(Payments$COLLECTION_DT)

#Create a new variable DATE and add it to the new data frame
payments3<-as.data.frame(mutate(payments2, DATE= (as.Date(payments2$COLLECTION_DT,
                                                          format = "%m/%d/%Y"))))

# arrange the date in order
payments3 <- payments3%>% arrange(payments3$DATE); head(payments3)
```

```
##    CASE_NUM PYMNT_AMT      COLLECTION_DT PYMNT_SRC PYMNT_TYPE   AP_ID
## 1 381388123     50.00   7/6/2002 0:00:00         A          C 1770778
## 2 381388123    160.00   8/5/2002 0:00:00         A          A 1770778
## 3 181384171      1.34  3/18/2005 0:00:00         A          C 1754083
## 4 121384407      0.72 10/27/2006 0:00:00         A          A 1726406
## 5 121384407      0.61 10/27/2006 0:00:00         A          A 1726406
## 6 691448592     28.13 12/18/2006 0:00:00         A          A 1720261
##         DATE
## 1 2002-07-06
## 2 2002-08-05
## 3 2005-03-18
## 4 2006-10-27
## 5 2006-10-27
## 6 2006-12-18
```

```
# i
# range of dates of all payments
range(payments3$DATE)
```

```
## [1] "2002-07-06" "2016-11-04"
```

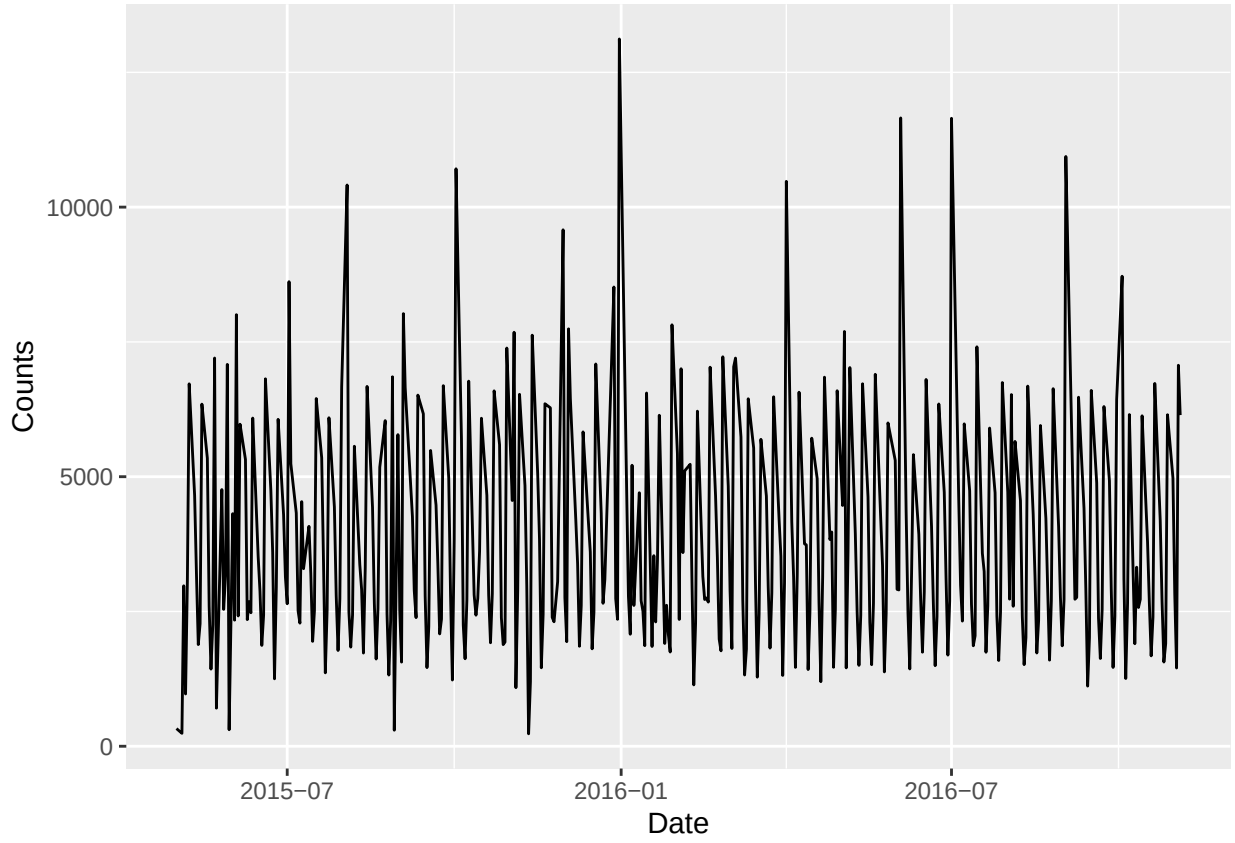
```
# ii
# the percentage of the total number of payments made before May 1, 2015
sum(payments3$DATE < "2015-05-01") / nrow(payments3)
```

```
## [1] 0.00381601
```

The range of dates of the data set go from 07-06-2002 to 11-04-2016 and the percentage of payments before May 5 2015 was 0.381601%.

3b) Show a sequence plot of the total number of payments made on each day from May 1, 2015 through the end of the data.

```
#Data set where all the payments were after May 1st, 2015
payments_over<-payments3[payments3$DATE>="2015-05-01", ]
payments_over <- payments_over%>% arrange(desc(payments_over$DATE))

#data frame with date and number of times a payment was made on that date
dates <- payments_over %>%
  group_by(DATE) %>%
  summarize(n = n())

ggplot(dates, aes(x=DATE, y=n)) +
  geom_line() +
  labs(x="Date", y="Counts")
```

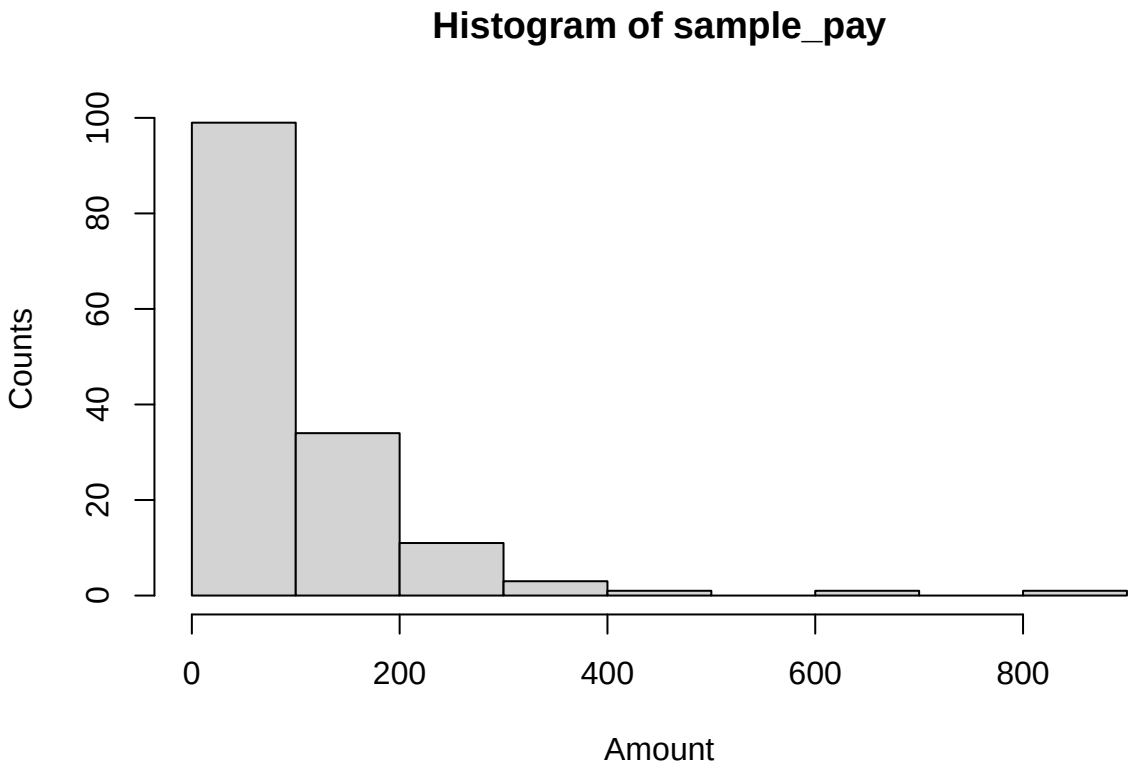


3c.

The bimodal shape of the marginal distribution of the number of payments over the time period can be explained by different billing cycles. The billing cycles usually spike near the end of every month and near the beginning of every month. For example, on the last day of 2015 we see the highest amount of payments made at 13119 payments. Another example of a spike in payments is on April 1st where 10478 payments were made. You can see that the payments spike towards the end of a month and near the beginning of every month.

3d.

```
set.seed(100)

#took a sample of 150 payments from payments3
sample_pay<-payments3$PYMNT_AMT[sample(nrow(payments3),150)]

hist(sample_pay, xlab = "Amount", ylab = "Counts")
```



The distribution of the payment amounts are right skewed with a large peak from 0-200. An explanation for the shape of the histogram could be that most of the payments made are small monthly payments and the big outlier payments could be because of missed or less frequent payments.

4a) It has been conjectured that parents deemed responsible for more children are more likely to make either a larger number of payments or a larger total payment amount over this period. Is that true?

H0: There is no correlation between more children and a higher amount of payments or payment amounts. (rho = 0)

HA:There is a correlation between having more children and a higher amount of payments or payment amounts. (rho != 0)

```
#arranged counts by the parent id
case_count<-Cases%>%
 arrange(AP_ID)

#shows the AP_ID of the parent and how many children they have.
```

```r
children_count<-as.data.frame(table(case_count$AP_ID))
names(children_count)<-c("AP_ID","Counts")
children_count$AP_ID=as.numeric(as.character(children_count$AP_ID))

#joined the payments3 and parents data frame together by AP_ID
parents_info<-left_join(payments3,Parents, by = "AP_ID")

#Parents ID and the total sum of their payments
parents_sum<-parents_info%>%
  group_by(AP_ID)%>%
  summarize(payment_total=sum(PYMNT_AMT))

#Parents ID and total amount of payments made

parent_payment<- parents_info%>%
  group_by(AP_ID)%>%
  summarize(num_payments = n())


#data frame with parent id, total payment and total children.
cost_children<-left_join(parents_sum, children_count,by = "AP_ID" )

cost_children<-left_join(cost_children, parent_payment,by = "AP_ID" )

# correlation hypothesis tests
# Assumptions : Both variables are quantitative and the linear
cor.test(cost_children$Counts, cost_children$payment_total)
```

```
##
##  Pearson's product-moment correlation
##
## data:  cost_children$Counts and cost_children$payment_total
## t = 50.146, df = 28577, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2737012 0.2950137
## sample estimates:
##       cor
## 0.2843926
```

```r
cor.test(cost_children$Counts, cost_children$num_payments)
```

```
##
##  Pearson's product-moment correlation
##
## data:  cost_children$Counts and cost_children$num_payments
## t = 67.138, df = 28577, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3590513 0.3790803
## sample estimates:
##       cor
## 0.3691087
```

There is a moderate positive correlation between having more children and having a larger total payment amount. The confidence interval shows us that we are 95% confident that the correlation will be between 0.2737012 and 0.2950137. There is also a moderate positive correlation between number of payments made and the number of children you have, with a correlation value of 0.369. Using the linear regression model, we can see that the p value for the number of children is much smaller than 0.05 so we reject the null, and we can conclude that there is a significant relationship between number of children and payment amount and number of children with number of payments.

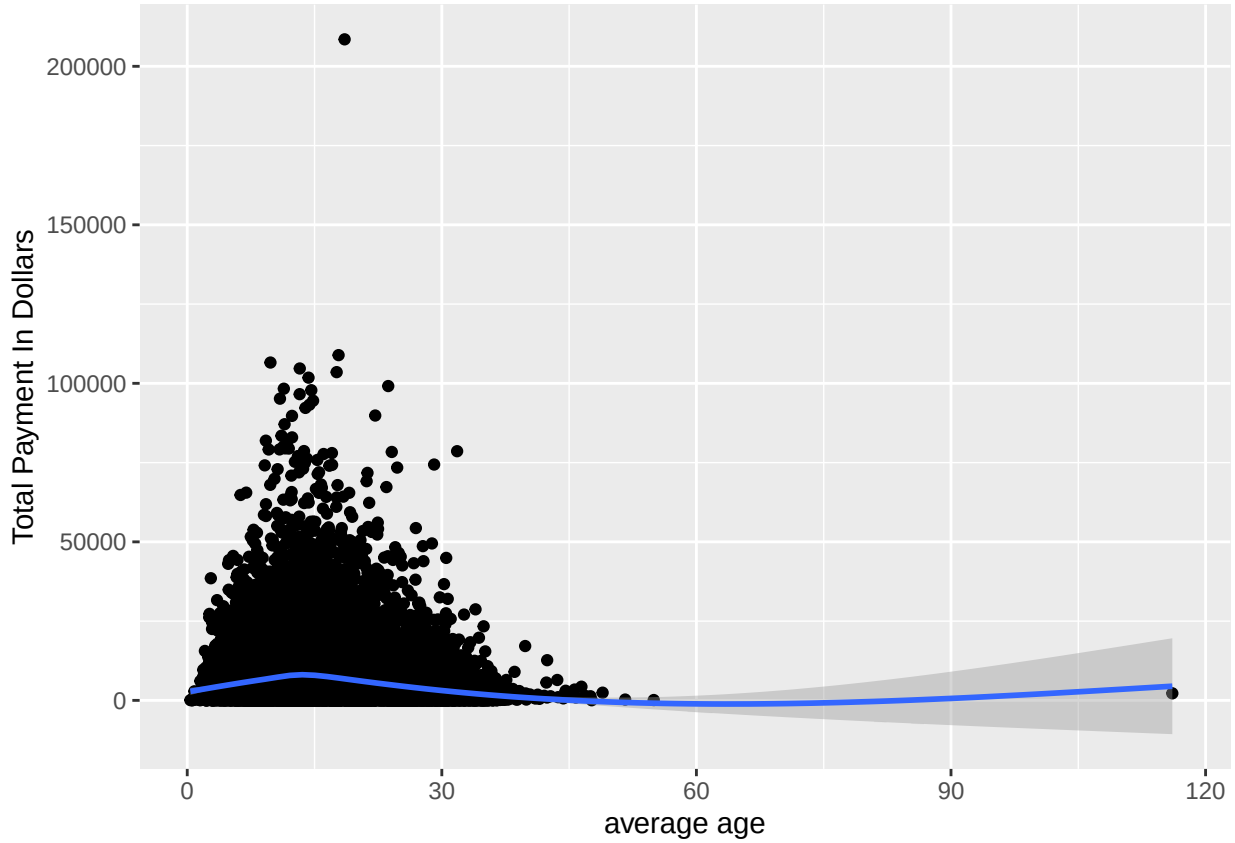
4b) It has been conjectured that parents responsible for younger children are more likely to make more payments. Is the average age of the children of an absent parent associated with the total amount of payments made by the absent parent?

H0 : There is no correlation or association between the average age of children of an absent parent and the total amount of payments made by the parent.

HA : There is an association between the average age of children of an absent parent and the total amount of payments made by the parent.

```
Children$DATE_OF_BIRTH_DT<-as.Date(Children$DATE_OF_BIRTH_DT, format="%m/%d/%Y")

# joined the parents, cases, and children data frames together by parent ID
# with new column AGE.
cases2<- left_join(Parents, Cases, by = "AP_ID")
cases2<-left_join(cases2, Children, by = "CASE_NUM")
cases2$age <- as.numeric(difftime(as.Date("2017-01-01"),
                                  cases2$DATE_OF_BIRTH_DT, units="days"))/365

#Average Age of kids per parent
avg_ages<-group_by(cases2,AP_ID)%>%
  summarize(mean_age= mean(age))
parents_avgage<-left_join(avg_ages,parents_sum, by="AP_ID", na.rm = TRUE)

ggplot(parents_avgage,aes(x=mean_age,y=payment_total))+
  geom_point()+
  geom_smooth()+
  labs(x = "average age", y = "Total Payment In Dollars")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 100946 rows containing non-finite values (stat_smooth).

## Warning: Removed 100946 rows containing missing values (geom_point).
```




```
# Assumptions passed
# Both variables are linear and quantitative so I can continue with the
# correlation hypothesis test
cor.test(parents_avgage$mean_age, parents_avgage$payment_total)
```

```
##
##  Pearson's product-moment correlation
##
## data:  parents_avgage$mean_age and parents_avgage$payment_total
## t = -20.551, df = 27369, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1349273 -0.1115935
## sample estimates:
##        cor
## -0.1232774
```

There is a weak negative correlation between average age of the children of an absent parent and the total payment amount. The confidence interval shows us that we are 95% confident that the correlation will be between -0.1349273 and -0.1115935.The P value smaller than 2.2e-16 indicate that it smaller than alpha 0.05. This mean I have sufficient evidence to reject the null hypothesis. From this, I can conclude that there is an association between the average age of children of an absent parent and the total amount of payments made by the parent.

4c) Does the location of the parent (AP_ADDR_ZIP) anticipate the total amount of payments made by the absent parent?

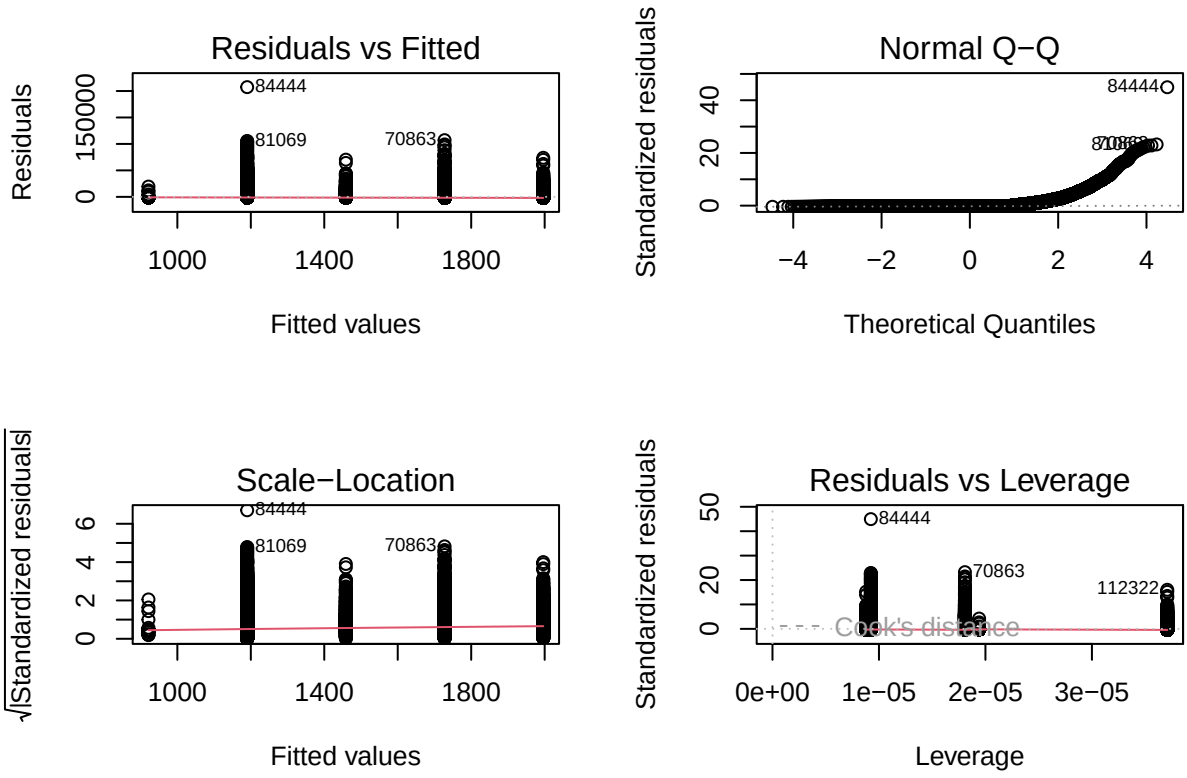
H0: there is no significant difference in the mean total amount of payments base on different locations of the parent

HA: there is significant difference in the mean total amount of payments base on different locations of the parent

```
#joined parents and the parents sum so that you can see the total payment amount per parent.
parents_adr<-left_join(Parents,parents_sum,by = "AP_ID")
parents_adr$AP_ADDR_ZIP<-as.integer(parents_adr$AP_ADDR_ZIP)
```

```
## Warning: NAs introduced by coercion
```

```
parents_adr[is.na(parents_adr)]<-0


#make dummy vars to do a model on the AP_ADDR_ZIP
dummy_vars <- model.matrix(~ factor(AP_ADDR_ZIP), data = parents_adr)
dummy_vars <- dummy_vars[, -1]
parents_adr <- cbind(parents_adr, dummy_vars)

parents_adr$AP_ADDR_ZIP_0 <- ifelse(parents_adr$AP_ADDR_ZIP == "0", 1, 0)

model <- lm(payment_total ~ AP_ADDR_ZIP, data = parents_adr)
par(mfrow = c(2,2))
plot(model)
```



```r
# From the plots, we can check the assumptions
# samples are independent and the variance are equal base on the residual vs fitted plot
# the response variable is approximately normal distributed base on the normal QQ plot
# since all assumption passed, we can conduct an one way anova test

summary(anova_model <- aov(payment_total ~ AP_ADDR_ZIP, data = parents_adr))
```

```
##                 Df    Sum Sq   Mean Sq F value Pr(>F)
## AP_ADDR_ZIP      1 1.485e+10 1.485e+10     698 <2e-16 ***
## Residuals   128315 2.730e+12 2.128e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p value from the test is smaller than 2e-16, which is smaller than alpha 0.05. I have sufficient evidence to reject the null hypothesis. This mean I can conclude that there is significant difference in the mean total amount of payments base on different locations of the parent

4d) Does the combination of attributes of the parent with the number and average age of the children involved predict the total amount of payments made by a parent?

Ho: All the beta equal to 0 (the predictor variables have no significant impact on the response variable)

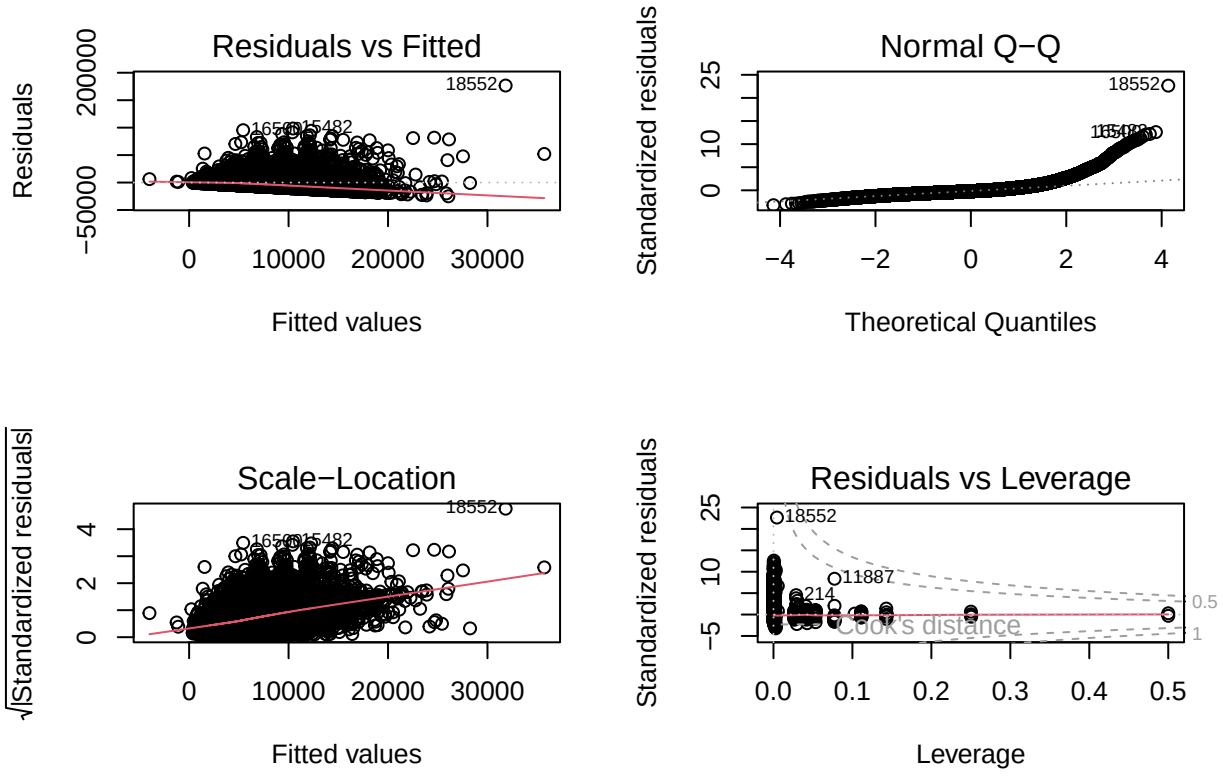
HA: Not all the beta equal to 0 (some of the predictor variables have significant impact on the response variable)

```r
library(dplyr)
library(tidyr)
library(ggplot2)
# A data frame containing the parents data frame and payment total, kid counts,
# and average kids age.

parents_kids2<-left_join(cost_children,avg_ages)
```

```
## Joining, by = "AP_ID"
```

```r
parents_kids3<-left_join(parents_kids2,Parents,by = "AP_ID")


#data set with relevant data, parent id, payment total, kid count,
# average kid age,marital status, primary language, and sex or parent

parents_kids_data<-parents_kids3[c(1,2,3,5,10,11,13)]
parents_kids_data[is.na(parents_kids_data)]<-0

#dummy vars for marital status, sex, and primary language
dummy_vars <- model.matrix(~ MARITAL_STS_CD, data = parents_kids_data)
dummy_vars<-dummy_vars[,-1]

dummy_vars2<-model.matrix(~SEX_CD, data= parents_kids_data)
dummy_vars2<-dummy_vars2[,-1]

dummy_vars3<-model.matrix(~PRIM_LANG_CD, data=parents_kids_data)
dummy_vars3<-dummy_vars3[,-1]
```

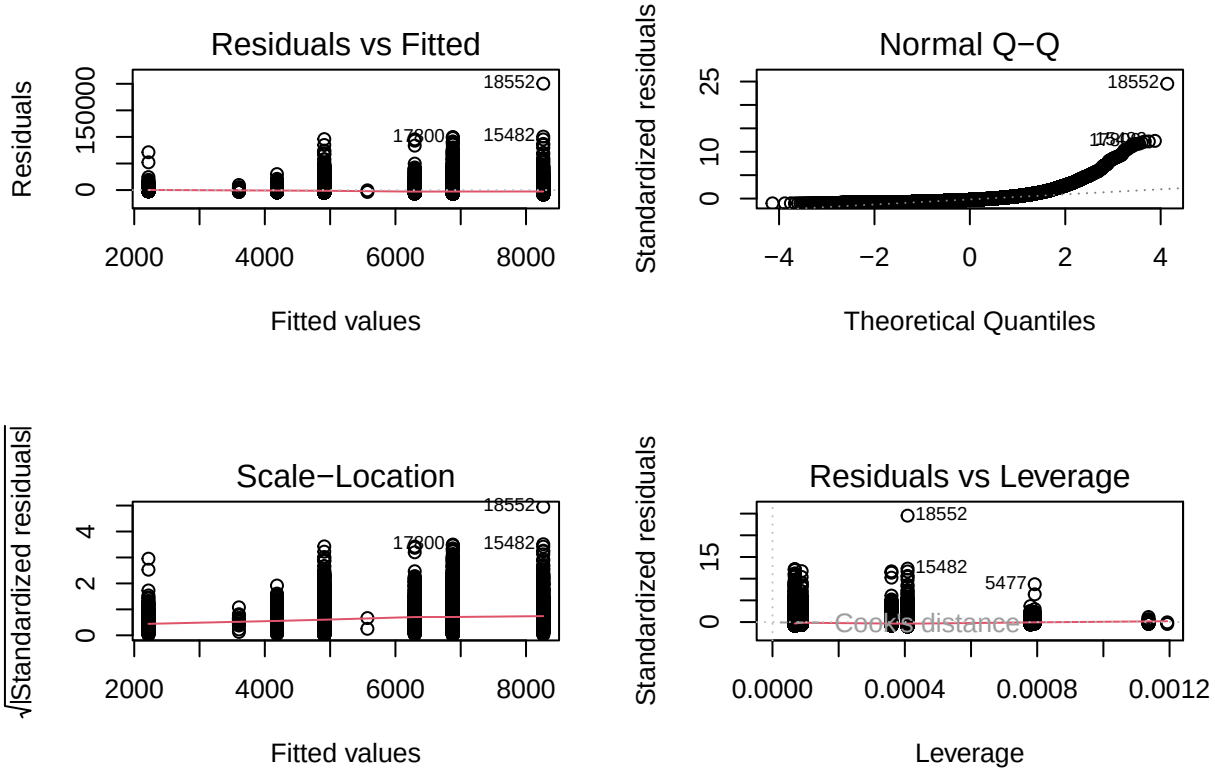
```r
#data from dummy vars that were left out
parents_kids_data <- parents_kids_data %>% mutate(SEX_CDF = ifelse(SEX_CD == "F", 1, 0))
parents_kids_data$SEX_CDF <- ifelse(parents_kids_data$SEX_CD == "F", 1, 0)

parents_kids_data <- parents_kids_data %>%
  mutate(PRIM_LANG_CD_blank = ifelse(PRIM_LANG_CD == " ", 1, 0))
parents_kids_data$PRIM_LANG_CD_blank <- ifelse(parents_kids_data$PRIM_LANG_CD == "", 1, 0)

parents_kids_data <- parents_kids_data %>%
  mutate(MARITAL_STS_CD_blank=ifelse(MARITAL_STS_CD == " ", 1, 0))


parents_kids_data <- cbind(parents_kids_data, dummy_vars,dummy_vars2,dummy_vars3)


model2 <- lm(payment_total ~ Counts + mean_age + MARITAL_STS_CD + SEX_CD +PRIM_LANG_CD,
             data = parents_kids_data)
par(mfrow = c(2,2))
plot(model2)
```




```r
# From the plots, we can check the assumptions
# samples are independent and the variance are equal base on the residual vs fitted plot
# the response variable is approximately normal distributed base on the normal QQ plot
# since all assumptions passed, I can continue on the global hypothesis test
summary(model2)
```

```
##
## Call:
## lm(formula = payment_total ~ Counts + mean_age + MARITAL_STS_CD +
##     SEX_CD + PRIM_LANG_CD, data = parents_kids_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25104  -4050  -1552   1886 176680
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2270.978    300.778   7.550 4.47e-14 ***
## Counts            2381.469     48.527  49.075  < 2e-16 ***
## mean_age           -81.156      6.018 -13.485  < 2e-16 ***
## MARITAL_STS_CDD   2538.218    263.728   9.624  < 2e-16 ***
## MARITAL_STS_CDM   1387.110    299.013   4.639 3.52e-06 ***
## MARITAL_STS_CDN    342.892    109.686   3.126  0.00177 **
## MARITAL_STS_CDS   2870.435    218.088  13.162  < 2e-16 ***
## MARITAL_STS_CDW   -579.827   1597.210  -0.363  0.71659
## SEX_CDM           2136.691    220.136   9.706  < 2e-16 ***
## SEX_CDU           4140.661   1379.183   3.002  0.00268 **
## PRIM_LANG_CDA    -3184.888   1800.031  -1.769  0.07685 .
## PRIM_LANG_CDE    -1355.921    155.433  -8.723  < 2e-16 ***
## PRIM_LANG_CDF     1081.743   2958.527   0.366  0.71464
## PRIM_LANG_CDH     -859.775   2609.796  -0.329  0.74182
## PRIM_LANG_CDI    -2386.292   1570.173  -1.520  0.12858
## PRIM_LANG_CDL    -6798.037   5528.686  -1.230  0.21886
## PRIM_LANG_CDO    -1539.433   1329.377  -1.158  0.24687
## PRIM_LANG_CDR     5548.526   2175.310   2.551  0.01076 *
## PRIM_LANG_CDS    -1789.188   1277.201  -1.401  0.16126
## PRIM_LANG_CDX    -3699.581   3911.480  -0.946  0.34425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7816 on 28559 degrees of freedom
## Multiple R-squared:  0.1031, Adjusted R-squared:  0.1025
## F-statistic: 172.8 on 19 and 28559 DF,  p-value: < 2.2e-16
```

```
parents_kids_data<-cbind(parents_kids_data,dummy_vars,dummy_vars2)

#model of the factors left out after making dummy variables,
#including the female sex, and the NA values from marital status and primary language
l<-lm(payment_total~SEX_CDF+PRIM_LANG_CD_blank+MARITAL_STS_CD_blank, data = parents_kids_data)
par(mfrow = c(2,2))
plot(l)
```




```
# From the plots, we can check the assumptions
# samples are independent and the variance are equal base on the residual vs fitted plot
# the response variable is approximately normal distributed base on the normal QQ plot
# since all assumptions passed, I can continue on the global hypothesis test for
# the factor that got left out
summary(l)
```

```
##
## Call:
## lm(formula = payment_total ~ SEX_CDF + PRIM_LANG_CD_blank + MARITAL_STS_CD_blank,
##     data = parents_kids_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8261  -4467  -2248   1505 200239
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6878.96      67.06  102.58   <2e-16 ***
## SEX_CDF              -2692.80     229.50  -11.73   <2e-16 ***
## PRIM_LANG_CD_blank    1385.28     161.84    8.56   <2e-16 ***
## MARITAL_STS_CD_blank -1968.24      98.09  -20.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8167 on 28575 degrees of freedom
```

```
## Multiple R-squared:  0.02015,    Adjusted R-squared:  0.02005
## F-statistic: 195.9 on 3 and 28575 DF,  p-value: < 2.2e-16
```

From the 2 tests, I can obtain that p value from the F-statistic and both p values from both tests are smaller than 2.2e-16, which is smaller than 0.05. From this, I have sufficient evidence to reject the null hypothesis and conclude that not all the beta equal to 0. This means that some of the predictor variables have significant impact on the response variable which is the total ammount of payments made by a parent

From the model we can also see that, for each additional child, the payment total is estimated to increase by 2434.36 dollars and for each year of the child the payment total decreases by 88.20 dollars. The p values on counts and mean age are lower than 0.001 which means that they are statistically significant and so we can assume that there is an association between counts and therefore we can use them to predict the total payments.

For the marital status and sex of the parent and total payments the p value was lower than 0.05 and we can also assume that they are statistically significant and they may predict the total amount paid by a parent.

The p values of the primary language are mostly higher than 0.05 which means we can assume that the primary language of the parent is not very significant when compared to total payments and therefore it is not a good predictor.

There are two models made because after making the dummy variables for the different factors, the female sex, and the na values were left out and the data from those factors are still relevant to our test.

In the data frame I made all NA values 0 because if we omit the NA values we would be omitting a factor and many of your samples. Therefore making NA values = 0 lets us keep those na values in our plots to make conclusions on.

Question 5

a) Among all parents who made payments, is there any association between the SD of total daily payments and the average of total daily payments?

As I conduct the correlation between the standard deviation of the daily payment of each parents and the average daily payment of each parent, I did get some NA value for the standard deviation. This was because some parent only paid once, therefore the standard deviation of 1 payment is 0 or undefined which result in NA value. However, there were only 1347 NA values out of of 28579 values which is around 5% of the sample which is low to make an impact on the sample. As a result, I removed the NA values. After removing the NA values, I conducted a correlation hypothesis test to check whether there is a relationship between the SD of total daily payments and the average of total daily payments.

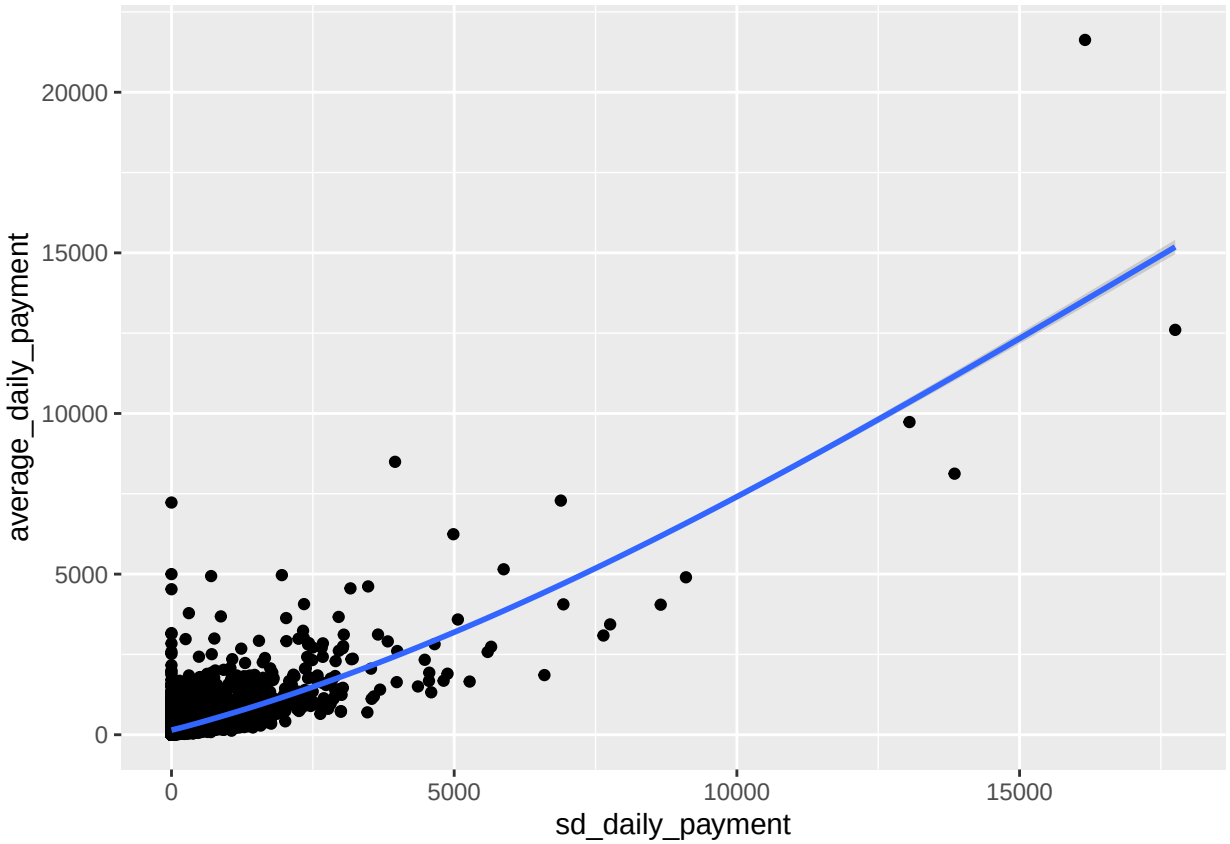
```
payments_from_parents <- Payments %>% group_by(AP_ID) %>%
  summarise(sd_daily_payment = sd(PYMNT_AMT),
            average_daily_payment = mean(PYMNT_AMT),
            total_nums_paydate = n_distinct(COLLECTION_DT))

# Omit the na in the standard deviation of the daily payment
payments_from_parents <- na.omit(payments_from_parents)
```

Null hypothesis : the correlation in the population is equal 0 (rho = 0)

Alternative hypothesis : the correlation in the population is not equal 0 (rho != 0)

```
ggplot(payments_from_parents,aes(x = sd_daily_payment, y = average_daily_payment))+
  geom_point()+
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
# Assumption Checking
# from the scatter plot below, the two variable is linear
# both variables are also quantitative
# Assumption passed, I can proceed with the correlation hypothesis test
cor.test(payments_from_parents$sd_daily_payment, payments_from_parents$average_daily_payment)
```

```
##
##  Pearson's product-moment correlation
##
## data:  payments_from_parents$sd_daily_payment and payments_from_parents$average_daily_payment
## t = 187.57, df = 27230, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7455779 0.7559425
## sample estimates:
##       cor
## 0.7508064
```

The p value here is less than 2.2e-16, which is smaller than alpha 0.05. From this, I have sufficient evidence to reject the null hypothesis. Since I reject the null hypothesis, I can conclude that the correlation in the population is not equal 0 (rho is not equal 0), therefore, there are association between the SD of total daily payments and the average of total daily payments. Additionally, the correlation value of 0.7508064 indicate that there is a strong positive relationship between these 2 variables.

b) The coefficient of variation (CV) is the ratio of the SD of daily payments to the mean. Show time sequence plots of the payments of 3 parents, with low, medium and high CV. That is, find three representative parents who make payments. One of these three should have a high CV, another an medium CV, and a third a low CV.

For this problem, I identify the range for the high, medium and low CV. Using the 25% and 75% quantile, I group all the CV values that are higher than 75% quantile to be considered as high CV, then I take the maximum value of that sample out to represent the parent that have a high CV. I do the same thing with low CV by grouping the values that lower than 25% quantile together then taking the maximum value out of the sample to represent that group. For the medium value, I group all the value that is lower than 75% quantile and higher than 25% quantile and take out the maximum value of that sample to represent the group. Then I filter the AP_ID that associate with that 3 CV's to obtain the collection date and payment amount information to construct the time sequence plot.

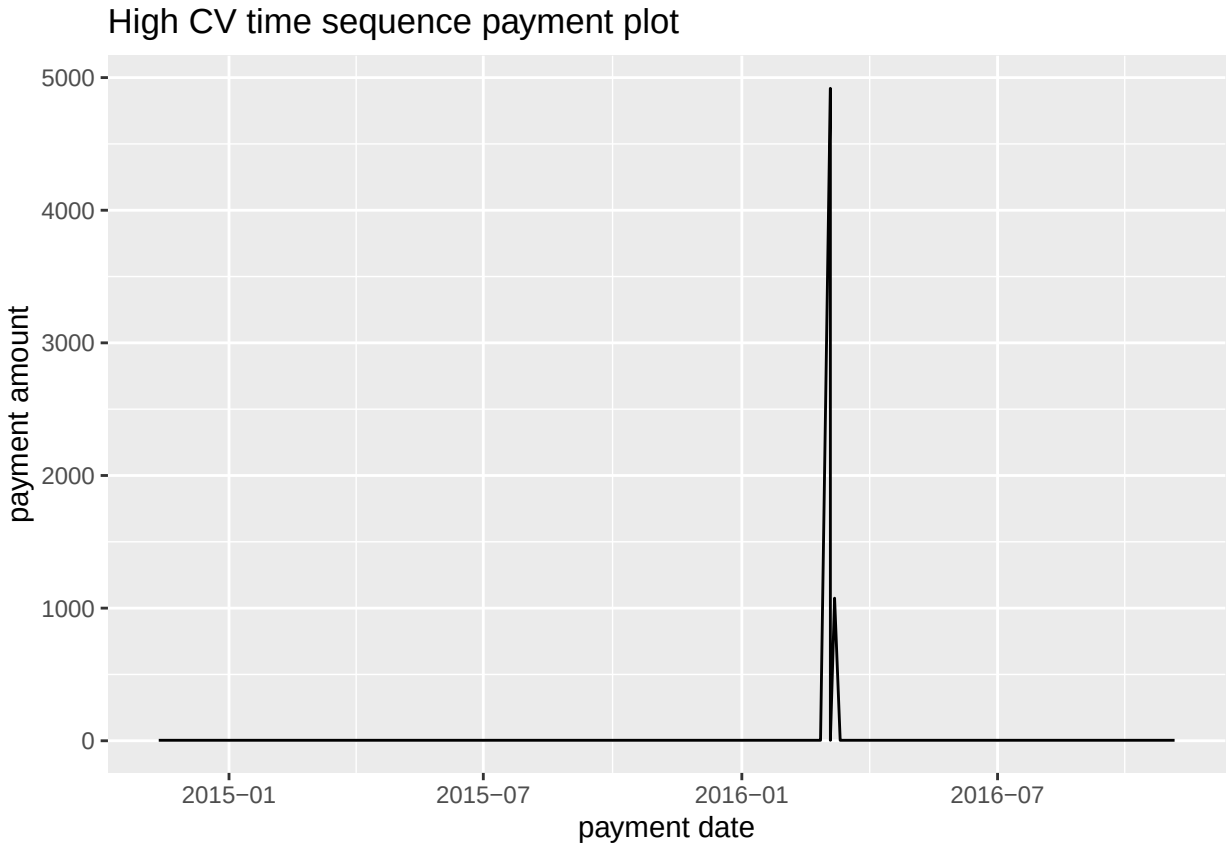
```
payments_from_parents <- payments_from_parents %>%
  mutate(CV_of_parents = sd_daily_payment / average_daily_payment) %>% arrange(desc(CV_of_parents))
```

```
high_cv <- max(payments_from_parents$CV_of_parents[payments_from_parents$CV_of_parents >
                                     quantile(payments_from_parents$CV_of_parents,
                                              prob = c(0.75))])

low_cv <- max(payments_from_parents$CV_of_parents[payments_from_parents$CV_of_parents <
                                     quantile(payments_from_parents$CV_of_parents,
                                              prob = c(0.25))])

medium_cv <- max(payments_from_parents$CV_of_parents[payments_from_parents$CV_of_parents >
                                                       low_cv &
                                                       payments_from_parents$CV_of_parents < high_cv])

payments_from_parents %>% filter(CV_of_parents == high_cv |
                                   CV_of_parents == low_cv |
                                   CV_of_parents == medium_cv)
```

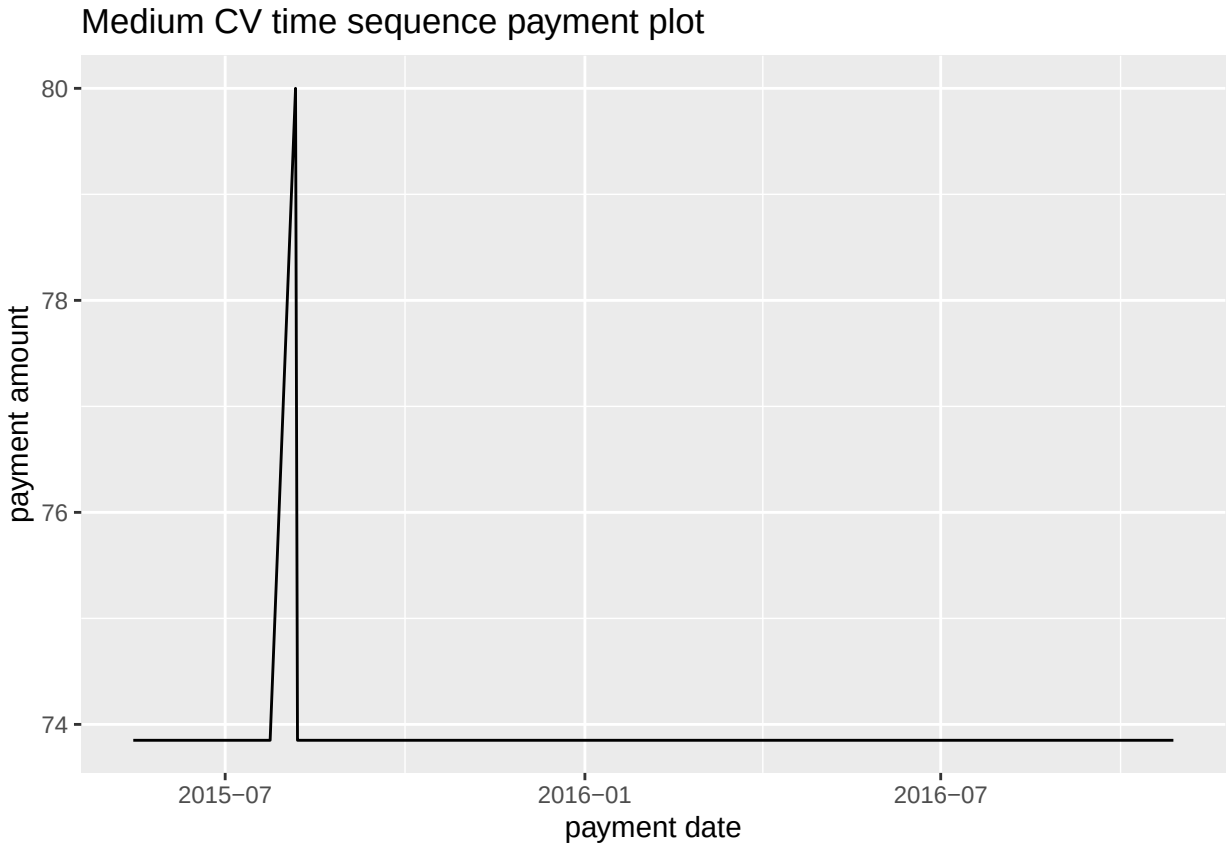
```
## # A tibble: 3 x 5
##     AP_ID sd_daily_payment average_daily_payment total_nums_paydate CV_of_pare~1
##     <int>            <dbl>                 <dbl>              <int>        <dbl>
## 1 1802326          380.                   37.7                  86      10.1
## 2 1754556          232.                   24.7                  77       9.38
## 3 1775478            0.972                74.0                  40       0.0131
## # ... with abbreviated variable name 1: CV_of_parents
```

Below here is the time sequence plots of the payments of 3 parents, with low, medium and high CV

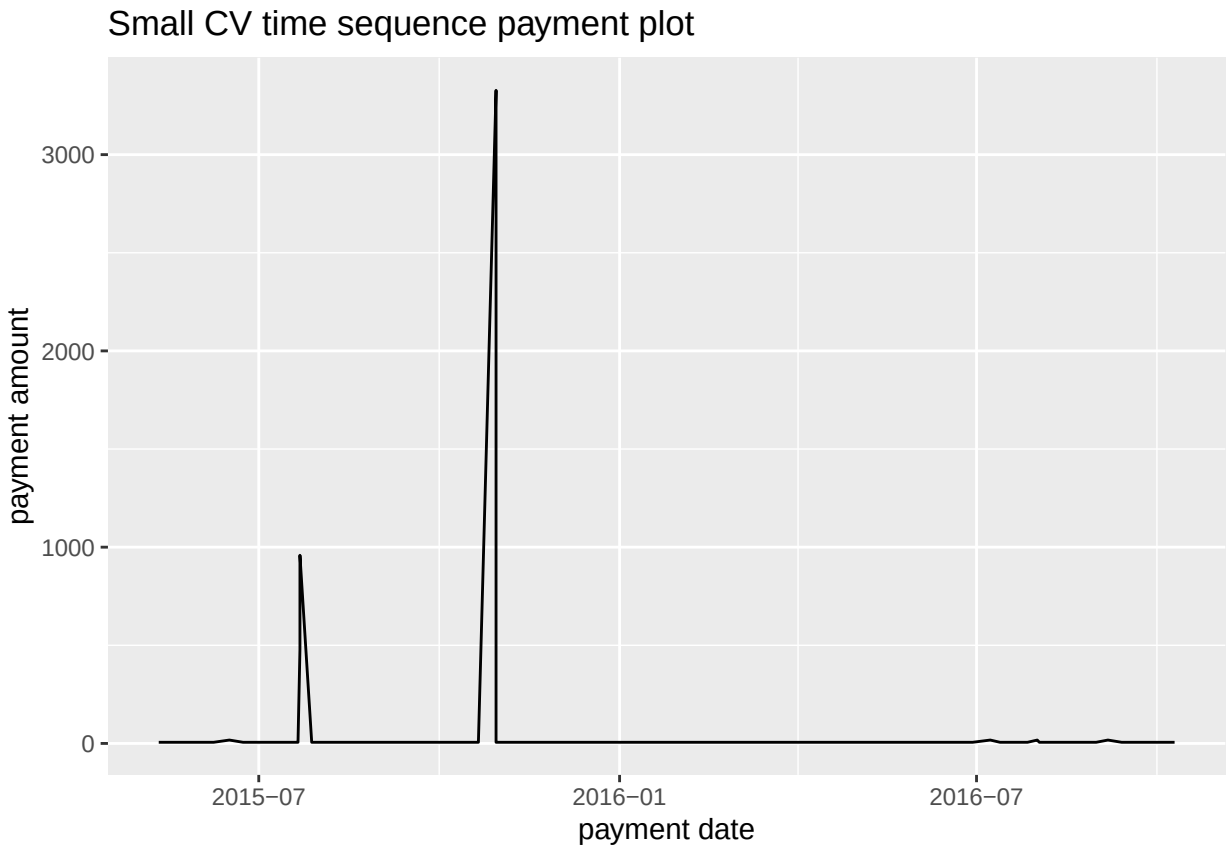
```
high_CV <- Payments %>% filter(AP_ID == "1802326") %>% group_by(COLLECTION_DT)

ggplot(high_CV, aes(x = as.Date(COLLECTION_DT, format = "%m/%d/%Y %H:%M:%S"), y = PYMNT_AMT))+
  geom_line()+
  labs(title = "High CV time sequence payment plot",
       x = "payment date",
       y = "payment amount")
```




```
medium_CV <- Payments %>% filter(AP_ID == "1775478") %>% group_by(COLLECTION_DT)

ggplot(medium_CV, aes(x = as.Date(COLLECTION_DT, format = "%m/%d/%Y %H:%M:%S"), y = PYMNT_AMT))+
  geom_line()+
  labs(title = "Medium CV time sequence payment plot",
       x = "payment date",
       y = "payment amount")
```




```
low_CV <- Payments %>% filter(AP_ID == "1754556") %>% group_by(COLLECTION_DT)

ggplot(low_CV, aes(x = as.Date(COLLECTION_DT, format = "%m/%d/%Y %H:%M:%S"), y = PYMNT_AMT))+
  geom_line()+
  labs(title = "Small CV time sequence payment plot",
       x = "payment date",
       y = "payment amount")
```



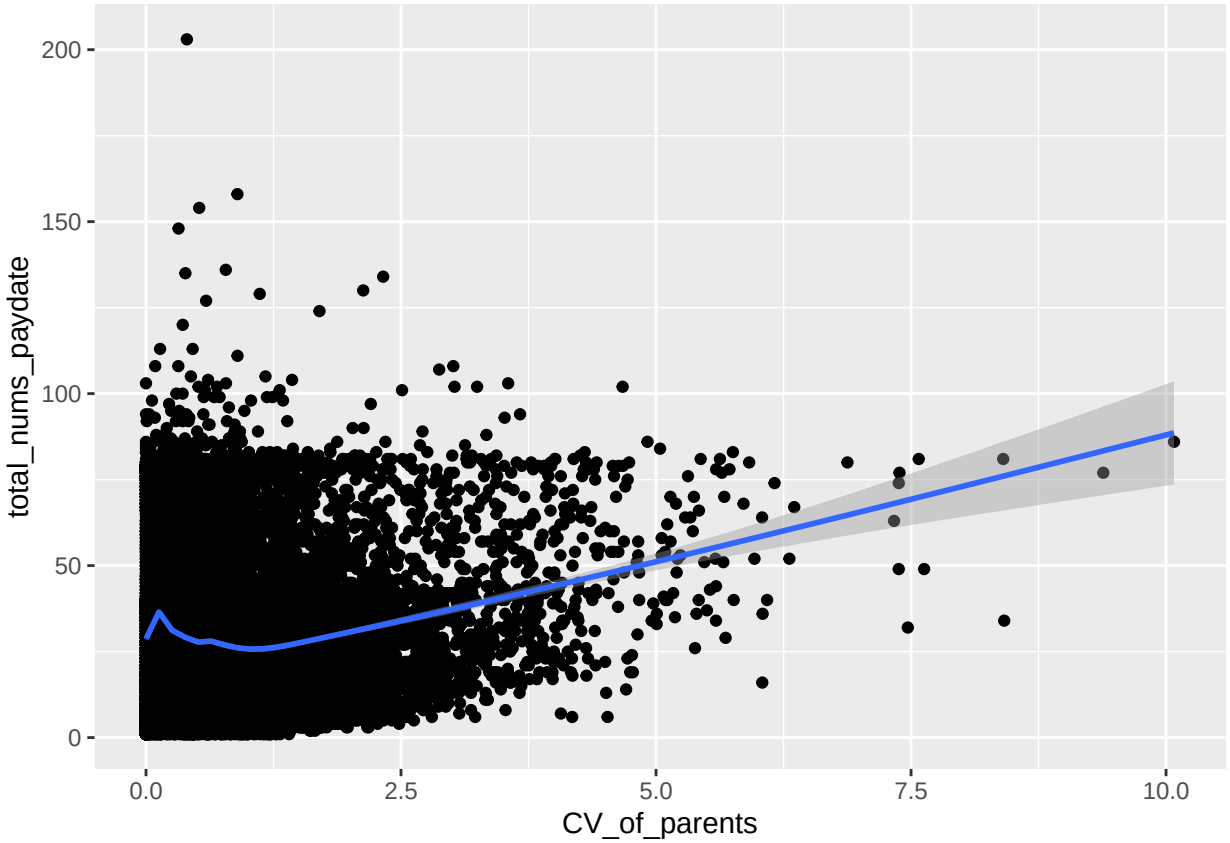
c) Is the CV of payments associated with the total amount of payments over this time period?

Null hypothesis : the correlation in the population is equal 0 (rho = 0)

Alternative hypothesis : the correlation in the population is not equal 0 (rho != 0)

```
ggplot(payments_from_parents, aes(x = CV_of_parents, y = total_nums_paydate))+
  geom_point()+
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
# Assumption Checking
# from the scatter plot below, the two variable is linear
# both variables are also quantitative
# Assumption passed, I can proceed with the correlation hypothesis test

cor.test(payments_from_parents$CV_of_parents, payments_from_parents$total_nums_paydate)
```

```
##
##  Pearson's product-moment correlation
##
## data:  payments_from_parents$CV_of_parents and payments_from_parents$total_nums_paydate
## t = 7.4745, df = 27230, p-value = 7.985e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03338999 0.05709564
## sample estimates:
##        cor
## 0.04524919
```

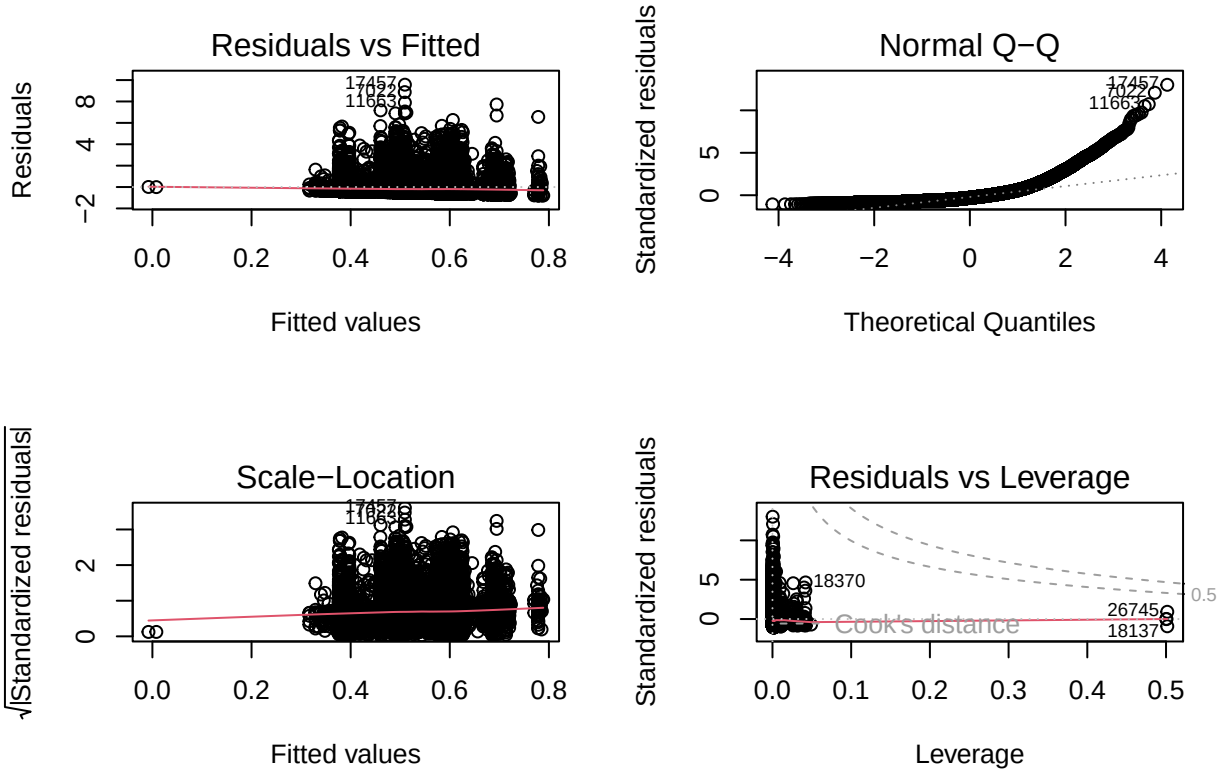
The p value here is 7.985e-14, which is smaller than alpha 0.05. From this, I have sufficient evidence to reject the null hypothesis. Since I reject the null hypothesis, I can conclude that the correlation in the population is not equal 0 (rho is not equal 0), therefore, there are association between the CV of payments with the total amount of payments over this time period. Additionally, the correlation is 0.04524919 which indicate that these 2 variables have a moderate correlation.

d) (Bonus Question) Do any attributes of the parent as revealed in these data anticipate that the parent will make consistent payments, that is, have small CV?

```
Combined_Parents_Payments <- Parents %>% inner_join(payments_from_parents, by = "AP_ID");

Combined_Parents_Payments <- Combined_Parents_Payments %>%
  select(-AP_CUR_INCAR_IND, -AP_DECEASED_IND) %>%
  mutate(AP_ZIP_Other = ifelse(AP_ADDR_ZIP == "na" |
                                 AP_ADDR_ZIP == "00" |
                                 AP_ADDR_ZIP == "04",1, 0),
         AP_ZIP_city = ifelse(AP_ADDR_ZIP == "01",1,0 ),
         AP_ZIP_South_state = ifelse(AP_ADDR_ZIP == "02",1,0 ),
         AP_ZIP_North_state = ifelse(AP_ADDR_ZIP == "03",1,0 ),
         MARITAL_STS_NA = ifelse(MARITAL_STS_CD == "",1,0 ),
         MARITAL_STS_Divorce = ifelse(MARITAL_STS_CD == "D",1,0 ),
         MARITAL_STS_Married = ifelse(MARITAL_STS_CD == "M",1,0 ),
         MARITAL_STS_N = ifelse(MARITAL_STS_CD == "N",1,0 ),
         MARITAL_STS_Single = ifelse(MARITAL_STS_CD == "S",1,0 ),
         MARITAL_STS_Widowed = ifelse(MARITAL_STS_CD == "W",1,0 ),
         Sex_Undefine = ifelse(SEX_CD == "U",1,0 ),
         Sex_Male = ifelse(SEX_CD == "M",1,0 ),
         Sex_Female = ifelse(SEX_CD == "F",1,0 ),
         Race_Asian = ifelse(RACE_CD == "A",1,0 ),
         Race_Black = ifelse(RACE_CD == "B",1,0 ),
         Race_Caucasion = ifelse(RACE_CD == "C",1,0 ),
         Race_Hispanic = ifelse(RACE_CD == "H",1,0 ),
         Race_N = ifelse(RACE_CD == "N",1,0 ),
         Race_P = ifelse(RACE_CD == "P",1,0 ),
         Race_Unknown = ifelse(RACE_CD == "U",1,0 ),
         Citizanship_Missing = ifelse(CITIZENSHIP_CD == "",1,0 ),
         Citizanship_Citizen = ifelse(CITIZENSHIP_CD == "C",1,0 ),
         Citizanship_Immigrant = ifelse(CITIZENSHIP_CD == "I",1,0 ),
         Citizanship_L = ifelse(CITIZENSHIP_CD == "L",1,0 ),
         Citizanship_R = ifelse(CITIZENSHIP_CD == "R",1,0 ),
         )
```

```
model3 <- lm(CV_of_parents~AP_ZIP_Other+
               AP_ZIP_city+
               AP_ZIP_South_state+
               MARITAL_STS_Divorce+
               MARITAL_STS_Married+
               MARITAL_STS_N+
               MARITAL_STS_Single+
               MARITAL_STS_Widowed+
               Sex_Male+
               Sex_Female+
               Race_Asian+
               Race_Black+
               Race_Caucasion+
               Race_Hispanic+
               Race_N+
               Race_P+
               Citizanship_Missing+
               Citizanship_Citizen+
```

```
              Citizanship_Immigrant+
              Citizanship_L+
              AP_APPROX_AGE, data =Combined_Parents_Payments)
par(mfrow = c(2,2))
plot(model3)
```




```
# From the plots, we can check the assumptions
# samples are independent and the variance are equal base on the residual vs fitted plot
# the response variable is approximately normal distributed base on the normal QQ plot
# since all assumptions passed, I can continue on the global hypothesis test
summary(model3)
```

```
##
## Call:
## lm(formula = CV_of_parents ~ AP_ZIP_Other + AP_ZIP_city + AP_ZIP_South_state +
##     MARITAL_STS_Divorce + MARITAL_STS_Married + MARITAL_STS_N +
##     MARITAL_STS_Single + MARITAL_STS_Widowed + Sex_Male + Sex_Female +
##     Race_Asian + Race_Black + Race_Caucasion + Race_Hispanic +
##     Race_N + Race_P + Citizanship_Missing + Citizanship_Citizen +
##     Citizanship_Immigrant + Citizanship_L + AP_APPROX_AGE, data = Combined_Parents_Payments)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7886 -0.4669 -0.2219  0.1672  9.5684
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.5463228  0.5365785   1.018 0.308611
## AP_ZIP_Other            0.0649724  0.0161614   4.020 5.83e-05 ***
## AP_ZIP_city            -0.0190541  0.0105849  -1.800 0.071853 .
## AP_ZIP_South_state     -0.0022722  0.0315751  -0.072 0.942633
## MARITAL_STS_Divorce     0.0141833  0.0254217   0.558 0.576902
## MARITAL_STS_Married     0.1025712  0.0287218   3.571 0.000356 ***
## MARITAL_STS_N           0.1106171  0.0101820  10.864  < 2e-16 ***
## MARITAL_STS_Single      0.0421133  0.0209954   2.006 0.044884 *
## MARITAL_STS_Widowed     0.0062659  0.1504066   0.042 0.966770
## Sex_Male                0.0381702  0.1284643   0.297 0.766372
## Sex_Female              0.1316701  0.1302754   1.011 0.312166
## Race_Asian             -0.0303867  0.1195419  -0.254 0.799348
## Race_Black             -0.0019500  0.0145435  -0.134 0.893342
## Race_Caucasion         -0.0022619  0.0230971  -0.098 0.921989
## Race_Hispanic          -0.0491879  0.0495357  -0.993 0.320729
## Race_N                 -0.0436099  0.1477945  -0.295 0.767942
## Race_P                 -0.5778972  0.5209401  -1.109 0.267295
## Citizanship_Missing    -0.1872047  0.5210085  -0.359 0.719364
## Citizanship_Citizen    -0.0730884  0.5209574  -0.140 0.888427
## Citizanship_Immigrant -0.1993922  0.5260245  -0.379 0.704650
## Citizanship_L          -0.0566828  0.5253063  -0.108 0.914072
## AP_APPROX_AGE           0.0001732  0.0003285   0.527 0.598150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7356 on 27210 degrees of freedom
## Multiple R-squared:  0.01098,    Adjusted R-squared:  0.01021
## F-statistic: 14.38 on 21 and 27210 DF,  p-value: < 2.2e-16
```

By conducting a global hypothesis test to determine if the model is significant with :

Null hypothesis : beta 1 = ... = beta k = 0

Alternate hypothesis : At least one of the beta is not equal to 0

Using the information above, the p value for the F distribution is smaller than 2.2e-16 which is smaller than alpha (0.05). Since the p value is smaller than alpha, I can reject the null hypothesis. Therefore, I have sufficient evidence to conclude that at least one of the beta is not equal to 0. This mean that the model is significant.

Next I will conduct a hypothesis test for each predictor variable to see find which predictors in this model is significant.

Null Hypothesis :

beta 1 = 0

. . .

. . .

. . .

beta 21 = 0

Alternative hypothesis :

beta 1 ! = 0

. . .

. . .

. . .

beta 21 != 0

```
# step 2
# get the p values that are smaller than alpha 0.05
as.list(summary(model)$coefficients[,4][summary(model)$coefficients[,4] < 0.05])
```

```
## $`(Intercept)`
## [1] 0
##
## $AP_ADDR_ZIP
## [1] 2.061291e-153
```

Step 3:

The attribute that have the p value that's smaller than 0.05 are : AP_ZIP_Other, MARITAL_STS_Married, MARITAL_STS_N, MARITAL_STS_Single

Step 4 :

The output above indicate that there are 4 betas that have a p value smaller than alpha 0.05. This mean I have sufficient evidence to reject the null hypothesis for AP_ZIP_Other, MARITAL_STS_Married, MARITAL_STS_N, MARITAL_STS_Single.

Step 5 :

Conclusion :

Since I have enough evidence to reject the null hypothesis AP_ZIP_Other = MARITAL_STS_Married = MARITAL_STS_N = MARITAL_STS_Single = 0, I can conclude that since the betas for these 4 variables is not 0, these attributes are most likely the attributes that anticipate that the parent will make consistent payments.