Assignment_1

Andrew Gutierrez

2022-09-11

1. Download a dataset from the web. You may use any source, but specify the source in your code. Also ensure that the data has a mix of quantitative and qualitative (categorical) variables.

For this assignment I chose to download a publicly-available dataset of vendor payments made by the City of Chicago from 1996 to present (although pre-2002 payments are uniformly listed with '2002' as the date, and data from 2003 to 2009 is also missing from the file). The City of Chicago has in recent years become a leader among US cities in making its civic data accessible to the public. This particular dataset contains both quantitative variables (ex. "Amount") and qualitative variables (ex. "Vendor Name" and "Department Name").

Link to download the data: https://data.cityofchicago.org/Administration-Finance/Payments/s4vu-giwb

2. Import the dataset into R

```
##### Note: the below file path may need to be adjusted, as it currently references a local location on my laptop
VP = read.csv("C:\\Users\\gutiera9\\Documents\\MSBA KSU\\CityofChicago_VendorPayments_1996toPresent.csv",header=T,sep=",")

head(VP)
```

```
##   VOUCHER.NUMBER   AMOUNT CHECK.DATE          DEPARTMENT.NAME CONTRACT.NUMBER
## 1  PV27212740089    16.00 12/28/2021                                       DV
## 2  PV38193801014   600.00   1/3/2020                                       DV
## 3  PV57215792314  3870.00 12/28/2021                                       DV
## 4  PV57215792296  2247.75 12/28/2021                                       DV
## 5 PVCI21CI103679 33854.96 11/29/2021 DEPT OF GENERAL SERVICES          12687
## 6 PVCI20CI204796   348.20  1/26/2021                                    33233
##                    VENDOR.NAME CASHED
## 1                  TERRY RIBAN  FALSE
## 2                       AMBIUS   TRUE
```

```
## 3              ELDER, JENISE C  FALSE
## 4              BOLINE, PATRICK  FALSE
## 5 SKYTECH ENTERPRISES, LIMITED   TRUE
## 6             OFFICE DEPOT INC   TRUE
```

3. Print out descriptive statistics for a selection of quantitative and categorical variables.

I'll first print out some descriptive statistics based on the quantitative "Amount" field. First, the sum total of all payments since 1996:

```
#### Sum total of all amounts
print(paste('$',sum(VP$AMOUNT),""))
```

```
## [1] "$ 80788910176.24 "
```

Next, the smallest and largest payments:

```
#### Minimum Amount
print(paste('Smallest amount paid to vendors: $',min(VP$AMOUNT[VP$AMOUNT>0]),""))
```

```
## [1] "Smallest amount paid to vendors: $ 0.01 "
```

```
#### Maximum Amount
print(paste('Largest amount paid to vendors: $',max(VP$AMOUNT), ""))
```

```
## [1] "Largest amount paid to vendors: $ 533413239.47 "
```

And finally, the average and median amounts paid out to vendors since 1996:

```
#### Average Amount
print(paste('Average amount paid to vendors: $',mean(VP$AMOUNT), ""))
```

```
## [1] "Average amount paid to vendors: $ 203583.639950811 "
```

```
#### Median Amount
print(paste('Median amount paid to vendors: $',median(VP$AMOUNT), ""))
```

```
## [1] "Median amount paid to vendors: $ 2495 "
```

Next, I'll move on to printing out descriptive statistics based on the categorical variables, starting with the number of vendors who have been paid by the City of Chicago since 1996:

```
#### Number of vendors paid:
n_distinct(VP$VENDOR.NAME)
```

```
## [1] 78998
```

Followed by the vendor who has received the most payments since 1996:

```
#### Use the Table function to create a subset of Vendor names and their frequencies
VP2 = as.data.frame(table(VP$VENDOR.NAME))

#### Find the maximum frequency from the subset
print(paste(max(VP2$Freq), "payments to", VP2$Var1[VP2$Freq == max(VP2$Freq)], " "))
```

```
## [1] "8149 payments to COMMONWEALTH. EDISON CO.  "
```

Now, the number of City of Chicago departments that have paid vendors since 1996:

```
#### Number of Departments that have paid vendors
n_distinct(VP$DEPARTMENT.NAME)
```

```
## [1] 56
```

And lastly, the Department that has made the most payments during that time:

```
#### Use the Table function to create a subset that contains only the dataframe rows that have a Department value listed
VP2 = as.data.frame(table(VP$DEPARTMENT.NAME[VP$DEPARTMENT.NAME != ""]))

#### Find the maximum frequency from the subset
print(paste(max(VP2$Freq), "payments made by", VP2$Var1[VP2$Freq == max(VP2$Freq)], " "))
```

```
## [1] "35195 payments made by DEPT OF FAMILY AND SUPPORT SERVICES  "
```

4. Transform at least one variable. It doesn't matter what the transformation is.

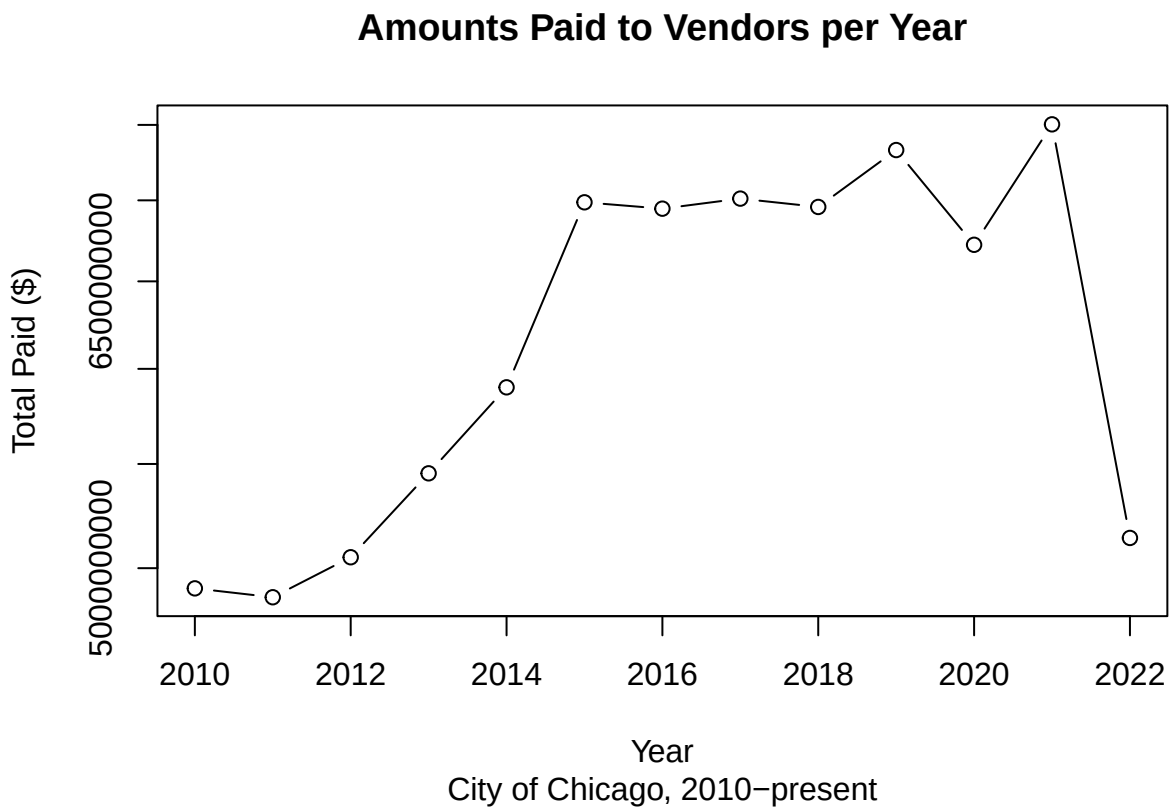
For this step, I'll do a quick transformation on the "Check Date" variable in order to extract the year the payment was made. First, since some of the date values listed contain only years (and not days or months), I'll apply some logic to fill in those incomplete date values with "January 1st". This will allow me to then run a conversion on all the values in the entire column in order to convert them from characters to dates, and then I'll create a new column with the "year" values from that date. I'll use this new column later for my plots.

```
#### fill in incomplete dates with "January 1st"
VP$YEAR = ifelse(nchar(VP$CHECK.DATE) < 5,paste('1/1/',VP$CHECK.DATE,sep=""),VP$CHECK.DATE)

#### Extract the year
VP$YEAR = format(as.Date(VP$YEAR, format="%m/%d/%Y"),"%Y")

#### Examine the dataframe with the new column
head(VP$YEAR)
```

```
## [1] "2021" "2020" "2021" "2021" "2021" "2021"
```

5. Plot at least one quantitative variable, and one scatterplot

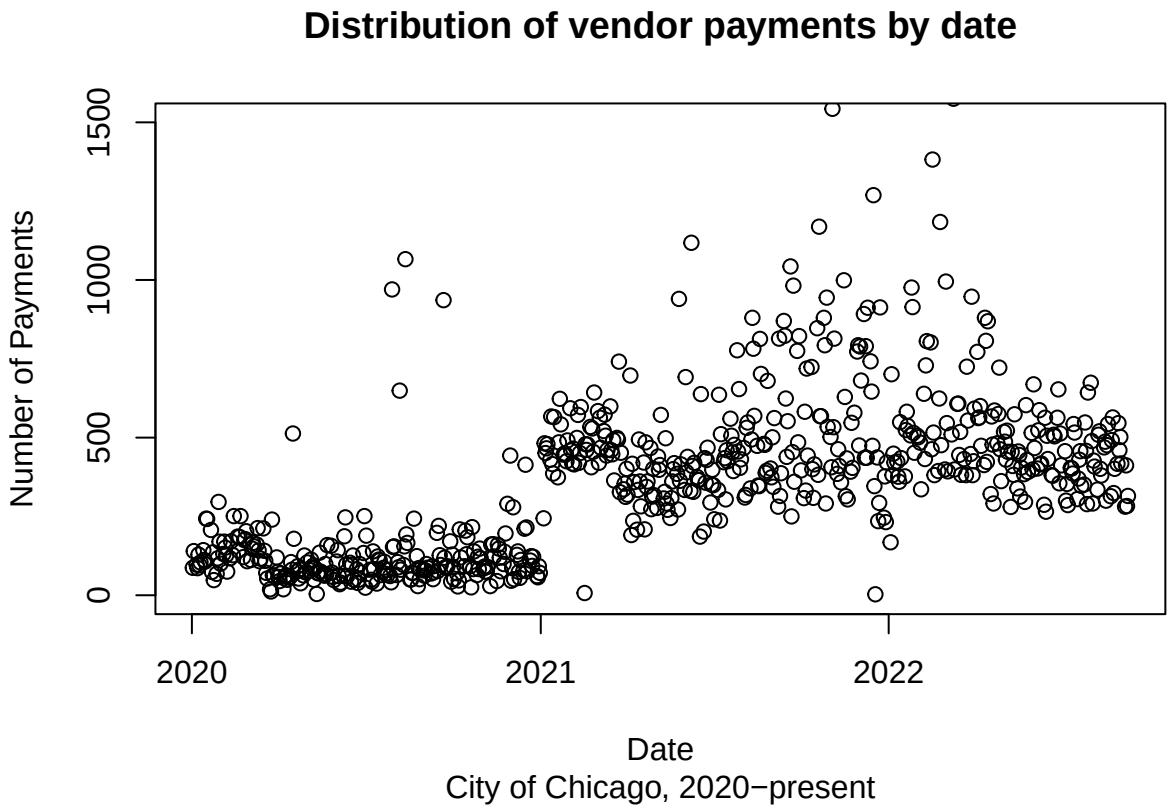
I'll first plot the distribution of amounts paid out from each year. Since pre-2002 data is all rolled up into one "bucket", for the sake of accuracy we'll exclude that data and only look at the years 2010-present (remember that 2003-2009 data is missing from the data source as well).



As you can see, this chart shows a steady rise in vendor payments from 2010 to 2015, followed by a period of flatness through 2019, somewhat of a drop in 2020 (perhaps due to the pandemic), and finally an increase in payments in 2021. 2022, of course, is still in progress.

Finally, I'll add a scatterplot that shows the distribution of the number of vendor payments made per day from 2020 to present (most pre-2020 dates are "incomplete" meaning they do not contain the full combination of year, month, and day; as such, I'll exclude pre-2020 dates from this plot):

```
#### Create subset based on aggregating the number of distinct payments made on each given day (excluding incomplete dates)
VP2 = aggregate(paste(VP$AMOUNT[nchar(VP$CHECK.DATE)>4],'-',VP$VENDOR.NAME[nchar(VP$CHECK.DATE)>4],""), list(as.Date(VP$CHECK.DATE[nchar(V

#### plot the subset
plot(VP2$Group.1, VP2$x,"p", main="Distribution of vendor payments by date",sub="City of Chicago, 2020-present",xlab="Date",ylab="Number o
```



The plot shows that the typical number of payments made per day during 2020 was quite static - ranging from 0 to 200. From 2021 to present though, the number of payments per day has been higher on average, and also far more varied: the distribution is notably more spread out - with some days having more than 1,000 payments - but it is generally centered around 500 payments per day.