Assignment 1.2

Emily Phillips

11/30/2021

Using the US Bureau of Labor Statistics data, choose a dataset that interests you. Then generate summary statistics for 2 variables, plot some of the features (e.g., histograms, box plots, density plots, etc.) of several variables, and save the data locally as CSV files.

My data source can be found at https://data.bls.gov/cgi-bin/surveymost. It represents data from 2011 to 2021 for monthly average hourly earnings of all employees, total private, and seasonally adjusted.

```
library(readxl)

#reading in dataset from excel file
avg_hourly <- read_excel("avg_hourly_earnings_all.xlsx")
head(avg_hourly)
```

```
## # A tibble: 6 x 13
##    Year   Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  2011  22.8  22.9  22.9  22.9  23.0  23.0  23.1  23.1  23.1  23.2  23.2  23.2
## 2  2012  23.2  23.3  23.4  23.4  23.4  23.5  23.5  23.5  23.6  23.6  23.6  23.7
## 3  2013  23.8  23.8  23.8  23.9  23.9  24.0  24.0  24.0  24.1  24.1  24.2  24.2
## 4  2014  24.2  24.3  24.3  24.3  24.4  24.4  24.5  24.6  24.6  24.6  24.6  24.6
## 5  2015  24.7  24.8  24.8  24.9  25.0  25.0  25.0  25.1  25.1  25.2  25.2  25.3
## 6  2016  25.4  25.4  25.4  25.5  25.6  25.6  25.7  25.7  25.8  25.9  25.9  25.9
```

```
#marking the Year column as a factor
avg_hourly$Year <- factor(avg_hourly$Year)
avg_hourly$Year
```

```
##  [1] 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
## Levels: 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
```

```
# mean,median,25th and 75th quartiles,min,max for Jan
print("Summary Statistics for 'Jan' variable")
```

```
## [1] "Summary Statistics for 'Jan' variable"
```

```
#handling error in dimnames with as.array
summ_Jan <- as.array(summary(avg_hourly$Jan))
print(summ_Jan)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.85   23.98   25.36   25.71   27.16   29.92
```

```
# mean,median,25th and 75th quartiles,min,max for Jun
cat("\nSummary Statistics for 'Jun' variable\n")
```

```
##
## Summary Statistics for 'Jun' variable
```

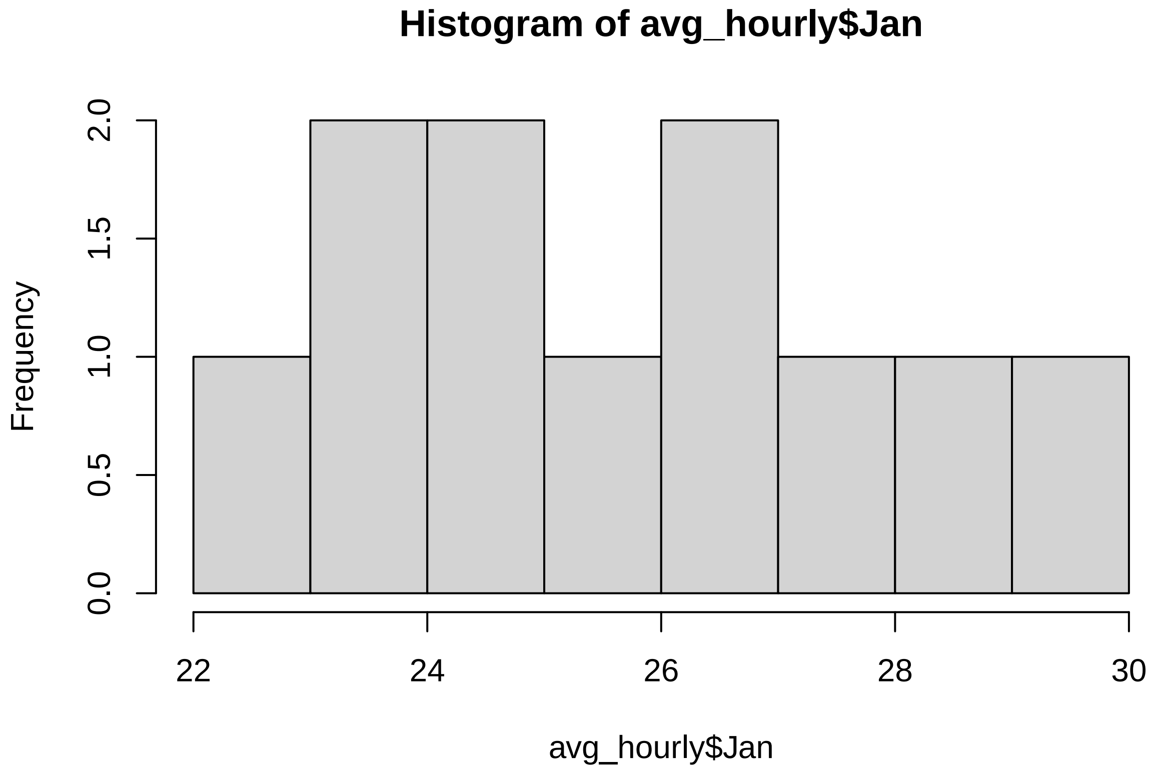
```
#handling error in dimnames
summ_Jun <- as.array(summary(avg_hourly$Jun))
print(summ_Jun)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.01   24.21   25.63   26.05   27.50   30.44
```

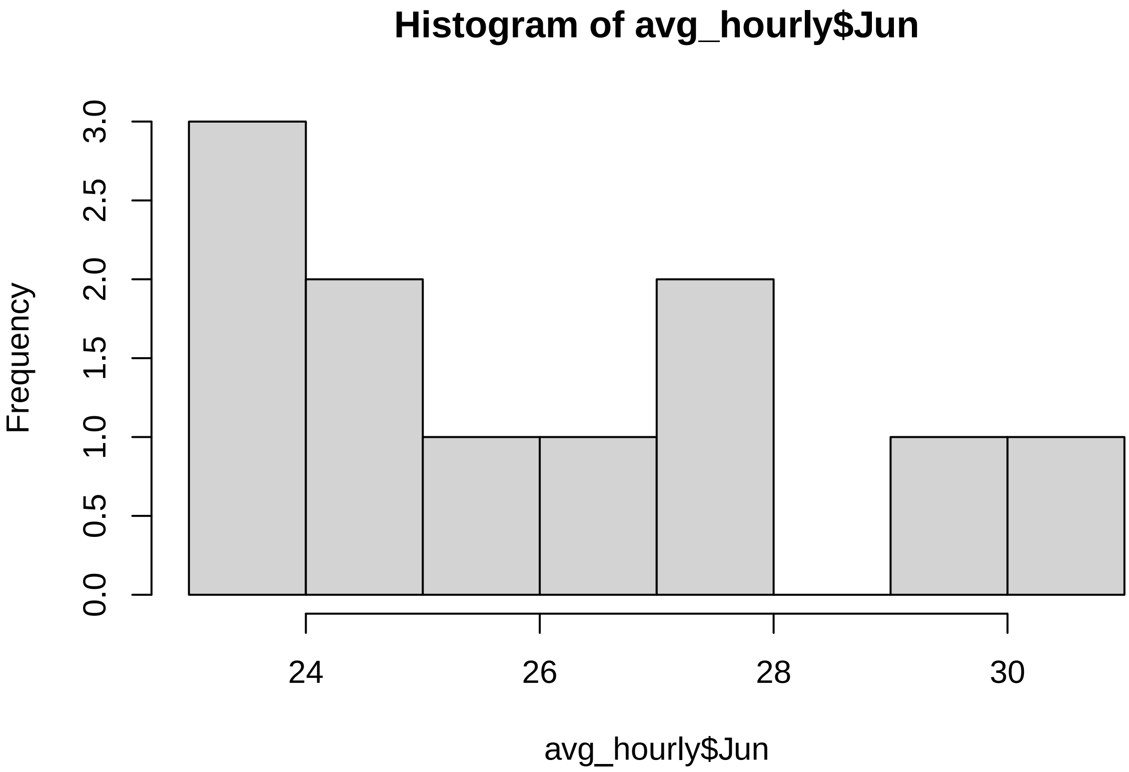
I found the summary statistics for two of the month variables: 'Jan' (January) and 'Jun' (June). From looking over the values, they have the almost the same median value, but are just slightly different in terms of the range of their numerical distributions. They also have slightly different mean values but as we'll see with their histograms below, as the distribution is not normal, we will look more at the medians rather than the means.

January has a slightly lower minimum and maximum than January; however, they mostly overlap which shows that they are not very different in what they represent. We can better visualize this with the box plots below!

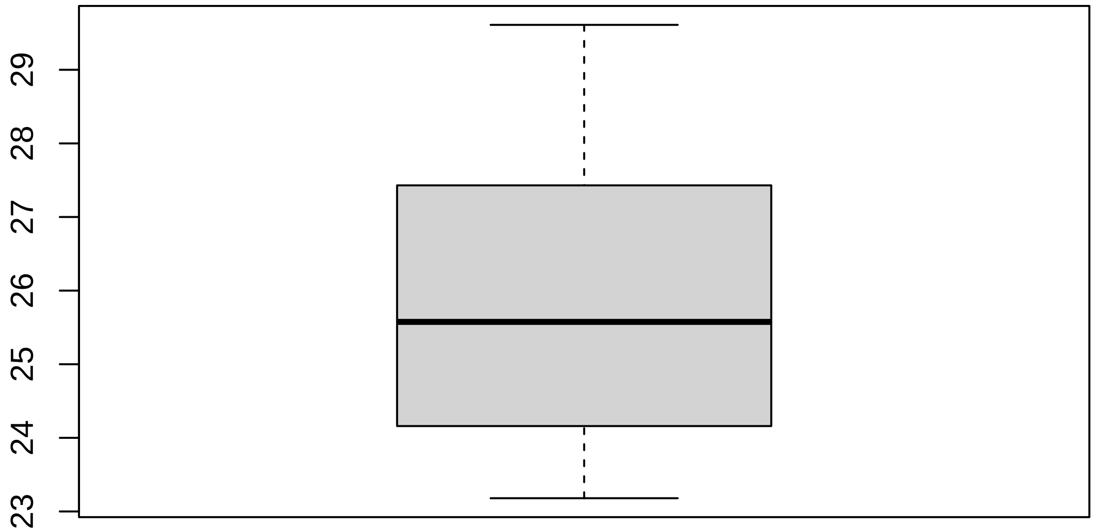
```
#histogram of few variables
#'Jan' variable as we have summary
hist(avg_hourly$Jan,breaks=6)
```



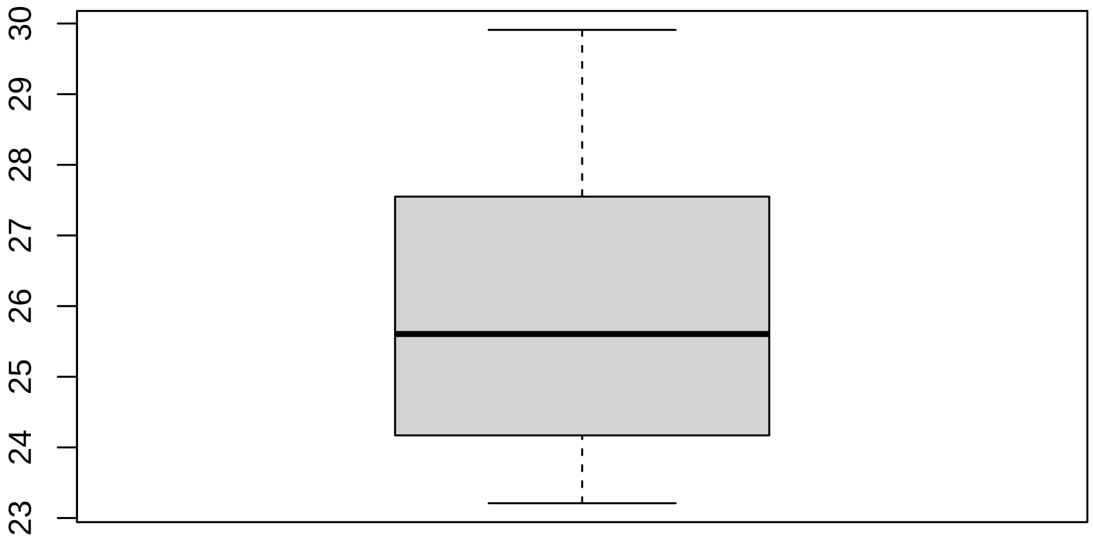
```
#histogram of 'Jun' variable
hist(avg_hourly$Jun,breaks=6)
```



```
#Distributions are more left-skewed --> more values at the lower end of the scale, lower average hourly
#Range only goes from 22.85 to 29.92, not a huge range but this could be standard for salaries dependin

#boxplot of other months
#November vs. December
boxplot(avg_hourly$Nov)
```



```
boxplot(avg_hourly$Dec)
```



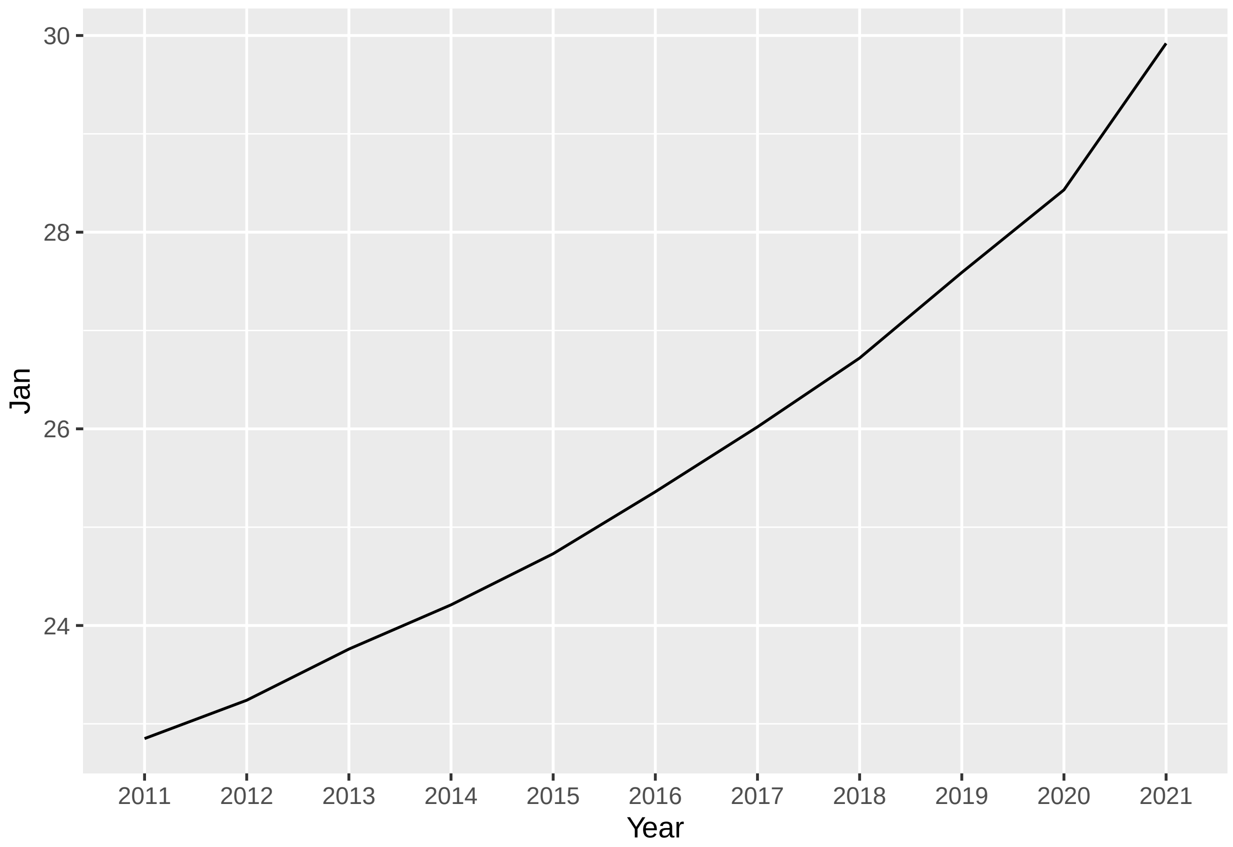
```
#Just slight difference in boxplots for November and December --> overlap in central 50% --> not a sign
```

Both histograms are mostly positive-skewed, with more values at the lower end of the distribution. This shows that a majority of employees have average hourly earnings around ~$22-$25. There is also a gap in the data in the 'Jun' distribution at about $28, showing that this earning is not well-represented for the private employees. As this distribution represents data from 2011-2021, we can also see that the majority of the years was spent at this lower-end; however, with some years, the average hourly earnings increased which could be due to growth for a company and how they performed for that given year.

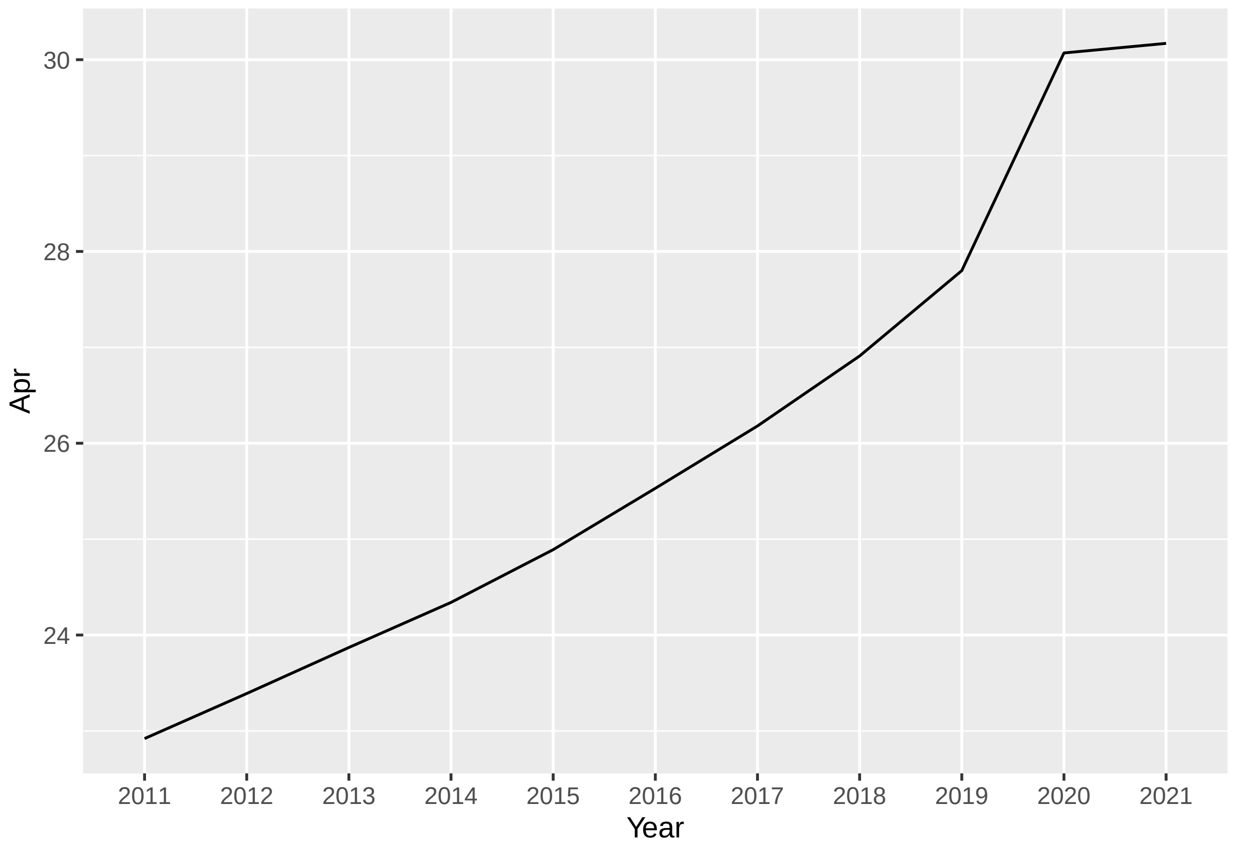
The histogram for January appears to neither be normal nor uniform. It has spikes, which could convey that transformations are needed on the variable. The distribution for June is not normal either. We cannot assume a normal distribution with these variables.

In terms of the boxplots, the variables mostly overlap in their spreads of distribution. As their medians are the same, it shows that the groups from November and December are not significantly different in their values of employees' average hourly earnings. There is not a significant difference between what people make in November compared to what they make in December, which makes sense since most promotions or changes in earnings happen towards the beginning or the end of a year. We might expect to see some sort of difference between January and December then (something to investigate). The boxplots also show a skewness towards the lower values in the range of the distribution, visualizing as well the positive skewness of the numerical variables.

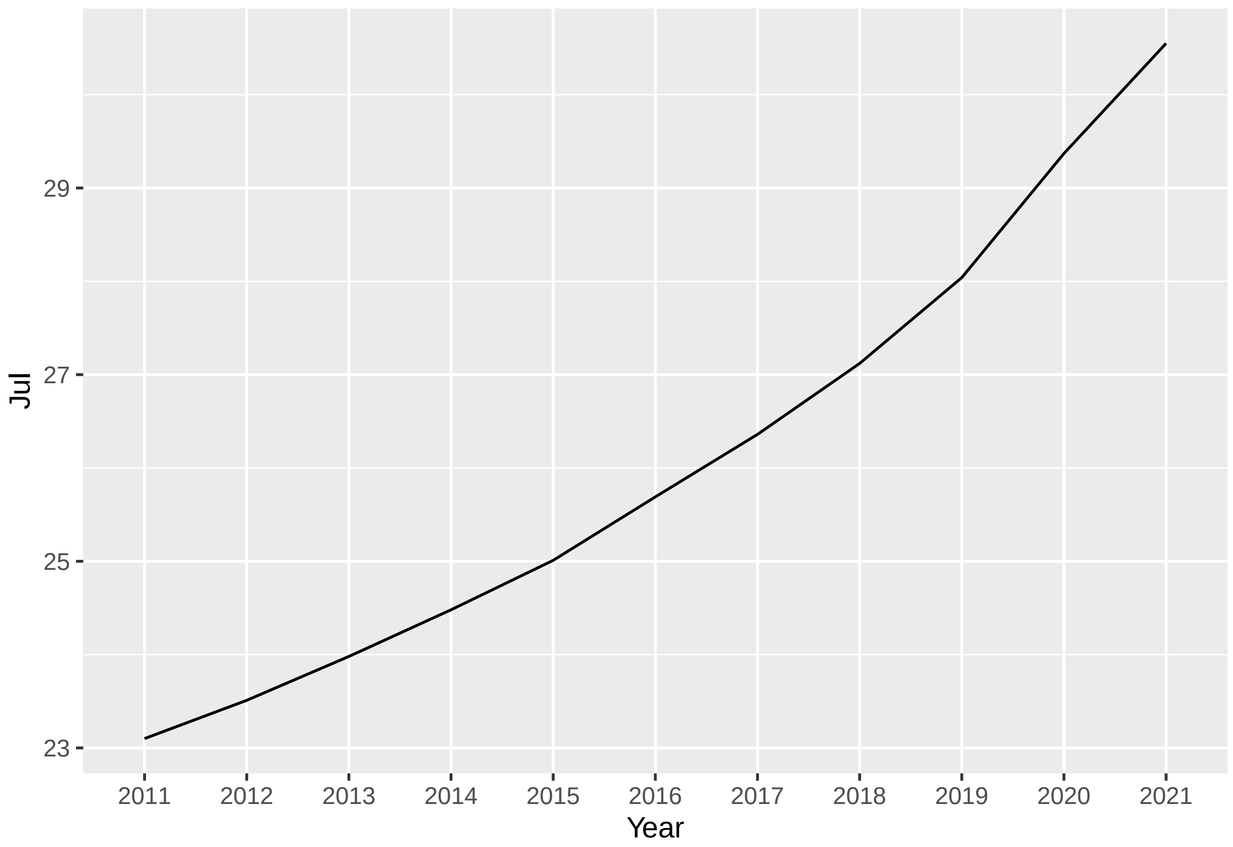
```
#line plot of one of the variables to show time-series
library(ggplot2)
ggplot(avg_hourly,aes(x=Year,y=Jan,group=1))+ geom_line()
```



```
ggplot(avg_hourly,aes(x=Year,y=Apr,group=1))+ geom_line()
```

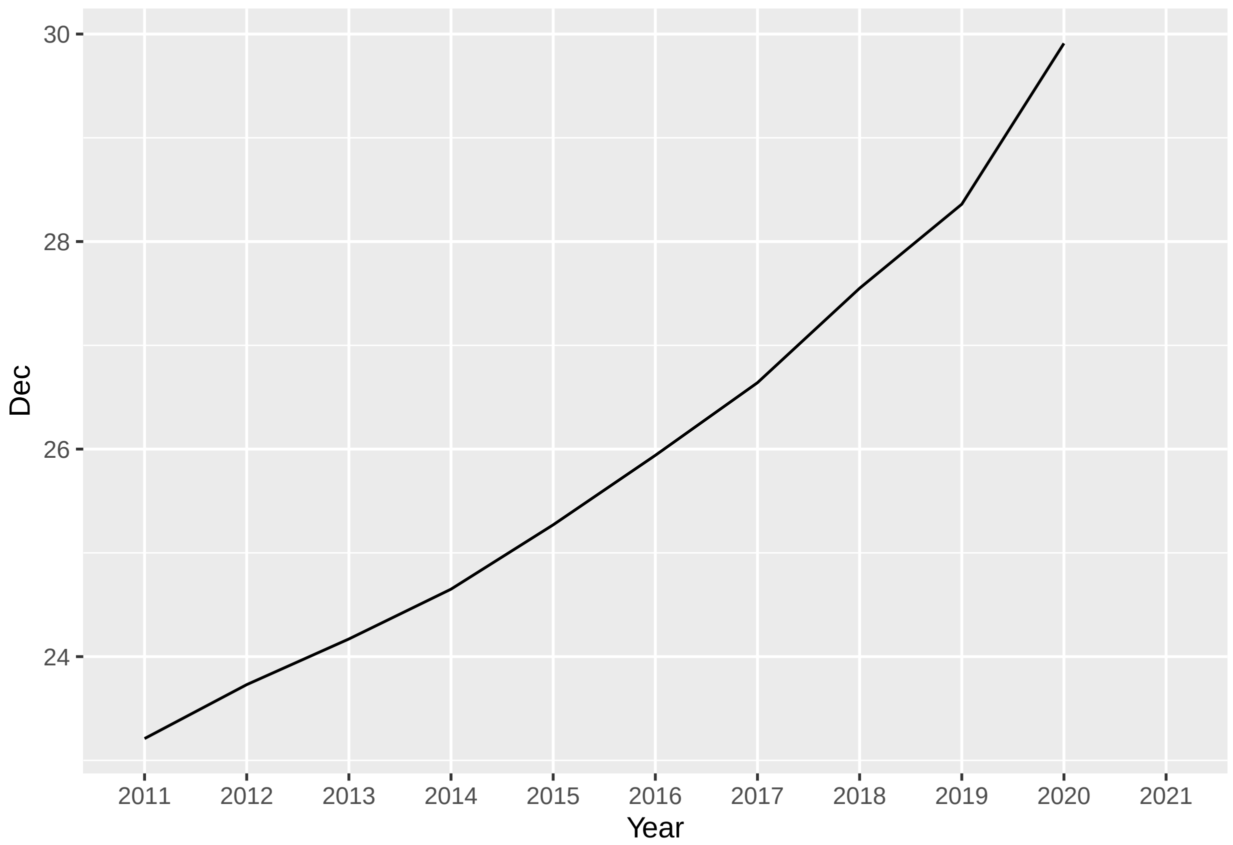


```
ggplot(avg_hourly,aes(x=Year,y=Jul,group=1))+ geom_line()
```



```
ggplot(avg_hourly,aes(x=Year,y=Dec,group=1))+ geom_line()
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



I wanted to look into one month from each quarter of the year to visualize the change over the years per month of total average hourly earnings for all private employees. From each of the line charts, it is clear that from 2011-2021 for each month, the total average hourly earnings experienced an increase. The increase was by about $8, which can represent a big change in overall salaray for employees! From the minimum of $22 hourly earnings which could be a ~45,760 yearly salary to $30 hourly earnings for a ~$62,400 salary, that can lead to quite a shift in one's lifestyle and is definitely money that matters. The months of April and July even experienced a maximum of earnings about $30, which can show some shift in those months that could be explored even more based on other data characteristics.

It is also difficult with the December data, as we don't have values yet for 2021! Therefore, we can only assume what will happen and with some extrapolation, we could possibly see that the average hourly earnings for all private employees will continue to increase in 2021 for the month.

In terms of my above observation as well with there being a significant difference between January and December earnings from year-to-year, from comparing the line charts, it seems that the slope for December increases at a slightly faster speed than the slope for January, potentially providing more clarity around the fact that the average hourly earnings will tend to increase at the end of the years with company bonuses, promotions, etc.

```r
#writing dataset to csv
write.csv(avg_hourly,'avg_hourly.csv')

#writing summary statistics to csv for January and June variables
write.csv(summ_Jan, file = 'summary_January.csv')

write.csv(summ_Jun, file = 'summary_June.csv')
```

Use the same dataset within the same website to explore some bivariate relations (e.g. bivariate plot, correlation, table cross table etc.)

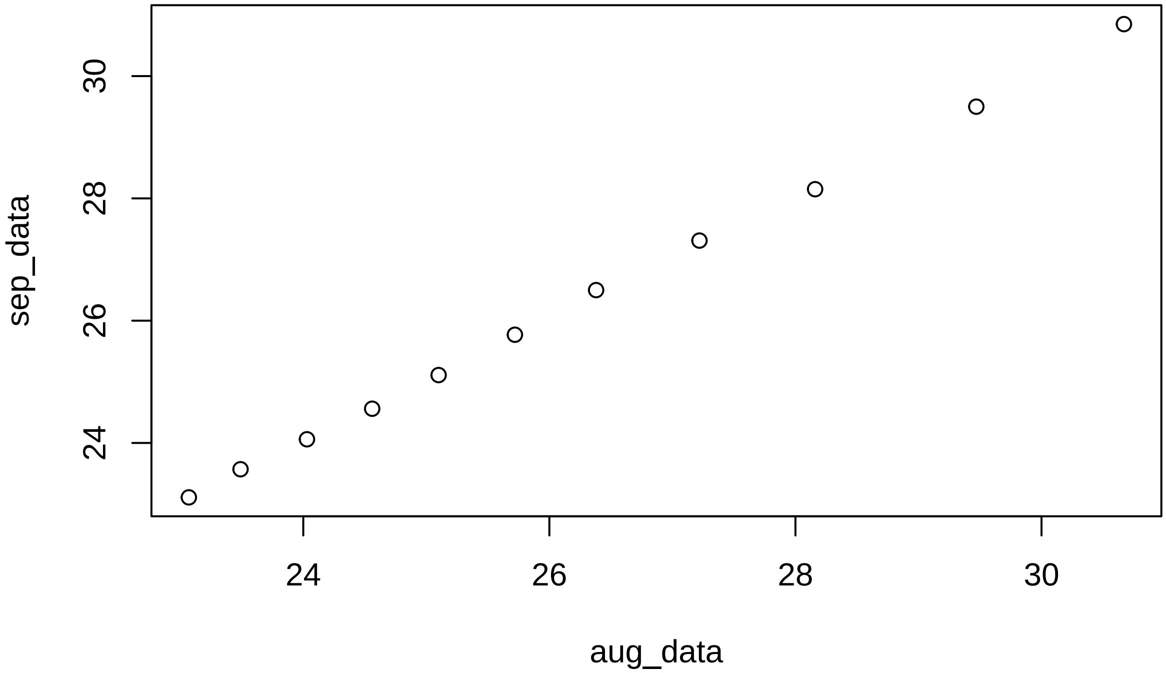
```
#August data
aug_data <- avg_hourly$Aug

#September data
sep_data <- avg_hourly$Sep

#correlation between August and September average hourly earnings from 2011-2016
cat("Correlation between August data and September data\n")
```

```
## Correlation between August data and September data
```

```
cor(aug_data,sep_data)
```

```
## [1] 0.9997838
```

```
#scatter plot to represent correlation, change in values
plot(aug_data,sep_data)
```



```
#correlation plot for the entire dataset
res <- cor(avg_hourly[,2:13])
round(res,2)
```

```
##      Jan  Feb  Mar  Apr  May Jun  Jul  Aug  Sep  Oct Nov Dec
## Jan 1.00 1.00 1.00 0.99 0.99   1 1.00 1.00 1.00 1.00  NA  NA
## Feb 1.00 1.00 1.00 0.99 0.99   1 1.00 1.00 1.00 1.00  NA  NA
## Mar 1.00 1.00 1.00 0.99 1.00   1 1.00 1.00 1.00 1.00  NA  NA
## Apr 0.99 0.99 0.99 1.00 1.00   1 0.99 0.99 0.99 0.99  NA  NA
## May 0.99 0.99 1.00 1.00 1.00   1 1.00 1.00 1.00 1.00  NA  NA
## Jun 1.00 1.00 1.00 1.00 1.00   1 1.00 1.00 1.00 1.00  NA  NA
## Jul 1.00 1.00 1.00 0.99 1.00   1 1.00 1.00 1.00 1.00  NA  NA
## Aug 1.00 1.00 1.00 0.99 1.00   1 1.00 1.00 1.00 1.00  NA  NA
## Sep 1.00 1.00 1.00 0.99 1.00   1 1.00 1.00 1.00 1.00  NA  NA
## Oct 1.00 1.00 1.00 0.99 1.00   1 1.00 1.00 1.00 1.00  NA  NA
## Nov   NA   NA   NA   NA   NA  NA   NA   NA   NA   NA   1  NA
## Dec   NA   NA   NA   NA   NA  NA   NA   NA   NA   NA  NA   1
```

When looking at the bivariate relations between August average hourly earnings from 2011-2021, and the September average hourly earnings, it turns out they have a correlation of 0.9998, which is essential 1.0 and a perfect positive correlation. This relationship is due to the fact that as the average hourly earnings in August increases year-to-year, the September average hourly earnings increases as well at basically the same race/pace. This high correlation though could indicate redundancy in the data, which should be considered. Also, correlation does not imply causation. Although we see that the average hourly earnings increase over the years for both months, it does not mean that there is a cause-and-effect relationship between the two.

The months year-to-year can differ immensely in the events that occur, and although it is good to see that the average hourly earnings increase which can show that things improve financially by 2021, it does not mean that August hourly earnings impact September hourly earnings. However, they could depending on various other factors that we'd have to look into!

I also created a correlation matrix for all of the numerical Month variables. Their correlations are extremely high and similar to what was stated above, perfect positive correlations. This could demonstrate redundancy in the data; however, because we are looking at time-series data, this can be understandable. There may not be much change in data from month-to-month. However, when we look at the changes from year to year, we see more of a significant difference and change over time. Therefore, while the months may be redundant in their average hourly earnings, the years really demonstrate where those earnings start to change and increase.

Generate a summary report. Make sure to include: summary for every variable, structure and type of data elements, discuss four results of your data.

```r
#summary for every variable in dataset
cat("Summary for every variable\n")
```

```
## Summary for every variable
```

```r
summary(avg_hourly)
```

```
##       Year        Jan             Feb             Mar             Apr
##  2011   :1   Min.   :22.85   Min.   :22.87   Min.   :22.87   Min.   :22.92
##  2012   :1   1st Qu.:23.98   1st Qu.:24.05   1st Qu.:24.06   1st Qu.:24.11
##  2013   :1   Median :25.36   Median :25.39   Median :25.45   Median :25.53
##  2014   :1   Mean   :25.71   Mean   :25.77   Mean   :25.82   Mean   :26.01
##  2015   :1   3rd Qu.:27.16   3rd Qu.:27.21   3rd Qu.:27.30   3rd Qu.:27.36
##  2016   :1   Max.   :29.92   Max.   :30.00   Max.   :29.97   Max.   :30.17
##  (Other):5
##       May             Jun             Jul             Aug
##  Min.   :22.99   Min.   :23.01   Min.   :23.10   Min.   :23.07  
##  1st Qu.:24.14   1st Qu.:24.21   1st Qu.:24.23   1st Qu.:24.30  
##  Median :25.57   Median :25.63   Median :25.69   Median :25.72  
##  Mean   :26.03   Mean   :26.05   Mean   :26.11   Mean   :26.17  
##  3rd Qu.:27.43   3rd Qu.:27.50   3rd Qu.:27.58   3rd Qu.:27.69  
##  Max.   :30.31   Max.   :30.44   Max.   :30.55   Max.   :30.67  
##                                                                 
##       Sep             Oct             Nov             Dec       
##  Min.   :23.11   Min.   :23.20   Min.   :23.18   Min.   :23.21  
##  1st Qu.:24.31   1st Qu.:24.34   1st Qu.:24.28   1st Qu.:24.29  
##  Median :25.77   Median :25.88   Median :25.57   Median :25.61  
##  Mean   :26.23   Mean   :26.28   Mean   :25.87   Mean   :25.94  
##  3rd Qu.:27.73   3rd Qu.:27.80   3rd Qu.:27.21   3rd Qu.:27.32  
##  Max.   :30.85   Max.   :30.96   Max.   :29.61   Max.   :29.91  
##                                  NA's   :1       NA's   :1      
```

```
#structure and type of data elements
cat("\nType of data elements\n")
```

```
## 
## Type of data elements
```

```
str(avg_hourly)
```

```
## tibble [11 x 13] (S3: tbl_df/tbl/data.frame)
##  $ Year: Factor w/ 11 levels "2011","2012",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Jan : num [1:11] 22.9 23.2 23.8 24.2 24.7 ...
##  $ Feb : num [1:11] 22.9 23.3 23.8 24.3 24.8 ...
##  $ Mar : num [1:11] 22.9 23.4 23.8 24.3 24.8 ...
##  $ Apr : num [1:11] 22.9 23.4 23.9 24.3 24.9 ...
##  $ May : num [1:11] 23 23.4 23.9 24.4 25 ...
##  $ Jun : num [1:11] 23 23.5 24 24.4 25 ...
##  $ Jul : num [1:11] 23.1 23.5 24 24.5 25 ...
##  $ Aug : num [1:11] 23.1 23.5 24 24.6 25.1 ...
##  $ Sep : num [1:11] 23.1 23.6 24.1 24.6 25.1 ...
##  $ Oct : num [1:11] 23.2 23.6 24.1 24.6 25.2 ...
##  $ Nov : num [1:11] 23.2 23.6 24.2 24.6 25.2 ...
##  $ Dec : num [1:11] 23.2 23.7 24.2 24.6 25.3 ...
```

Since the column 'Year' represents a limited number of values (2011-2021), I converted it from a numeric data type to a factor, so it can better represent the categories/groupings in the data. In terms of the summary for its variable, the values are 1 since the factor appears only once in the data set for each month of data. It represents a count of one. The other variables however, which represent the months in the year, are numerical as they convey the total average hourly earnings of all private employees (seasonally adjusted). From looking over the summary statistics, there really is not a significant difference between the range of values from 2011-2021, as mentioned above with the visualizations for single and bivariate relations. The average hourly earnings mostly lie between $22-$30, and there do not seem to be any significant outliers outside of this distribution. I think this consistency in the data from year-to-year shows how earnings don't vary very much for employees, especially at certain levels of an organization. Salaries are rather standardized especially within corporate organizations, so it makes sense that there is not a drastic difference from month-to-month. However, I think the maximums in the monthly values show how during certain years, there can be an increase in earnings based on certain factors like promotions, company pay-scale changes, etc. I think these can be good indicators of company performance over the years and how it impacts the earnings that employees make! The yearly changes depict more significance than the monthly changes, it seems. There are two missing values: one in the 'Nov' column and the other in the 'Dec'. As this is a small dataset of only 11 rows, one missing value could cause an impact in our modeling of the data. We would have to figure out how to handle these missing values when moving forward. The structure of the data is 11 x 13, 11 rows by 13 columns. The eleven rows go from years 2011-2021, and the columns are: Year, Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov and Dec. It is a rather small dataset, but since it represents time-series data, we wouldn't expect for the data to be too large, since we are just looking at data within a certain period of time. This data represents yearly data at the month perspective for those average hourly earnings of private employees.