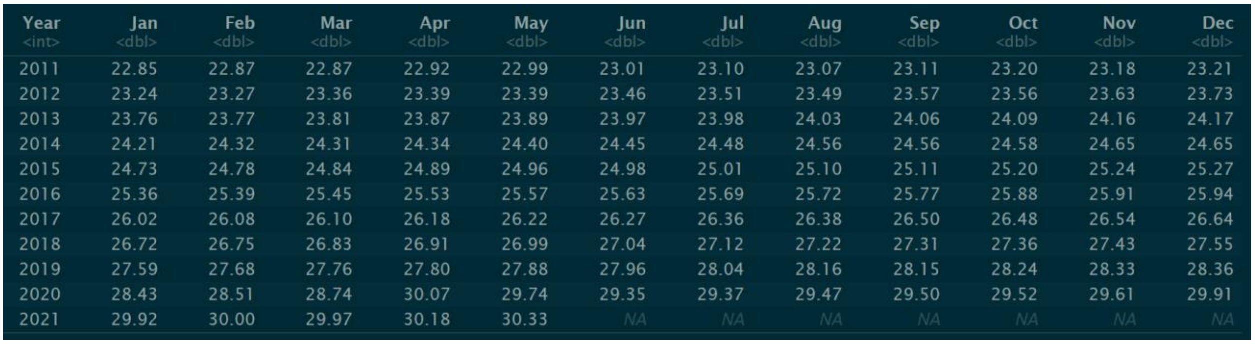
DSC 630

Laura Hoffmann

Assignment 1.2

Detailed Report

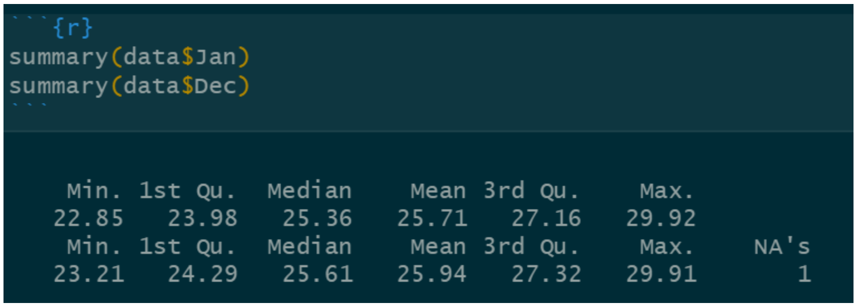
The Data



Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2011	22.85	22.87	22.87	22.92	22.99	23.01	23.10	23.07	23.11	23.20	23.18	23.21
2012	23.24	23.27	23.36	23.39	23.39	23.46	23.51	23.49	23.57	23.56	23.63	23.73
2013	23.76	23.77	23.81	23.87	23.89	23.97	23.98	24.03	24.06	24.09	24.16	24.17
2014	24.21	24.32	24.31	24.34	24.40	24.45	24.48	24.56	24.56	24.58	24.65	24.65
2015	24.73	24.78	24.84	24.89	24.96	24.98	25.01	25.10	25.11	25.20	25.24	25.27
2016	25.36	25.39	25.45	25.53	25.57	25.63	25.69	25.72	25.77	25.88	25.91	25.94
2017	26.02	26.08	26.10	26.18	26.22	26.27	26.36	26.38	26.50	26.48	26.54	26.64
2018	26.72	26.75	26.83	26.91	26.99	27.04	27.12	27.22	27.31	27.36	27.43	27.55
2019	27.59	27.68	27.76	27.80	27.88	27.96	28.04	28.16	28.15	28.24	28.33	28.36
2020	28.43	28.51	28.74	30.07	29.74	29.35	29.37	29.47	29.50	29.52	29.61	29.91
2021	29.92	30.00	29.97	30.18	30.33	NA	NA	NA	NA	NA	NA	NA

A small data set showing the average hourly wages in the years 2011-2021 with the months as the columns, (the variables).

Generate Summary Statistics for Two Variables

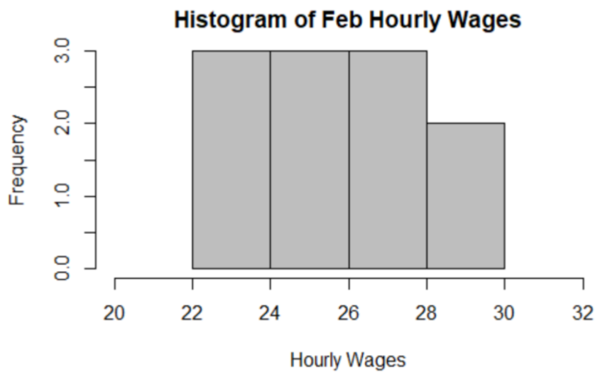


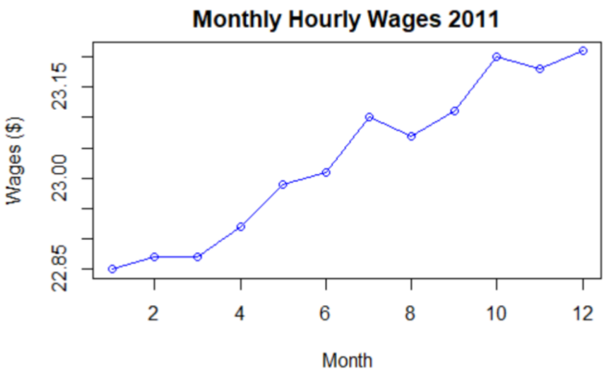
~~~
```{r}
summary(data$Jan)
summary(data$Dec)
```
~~~

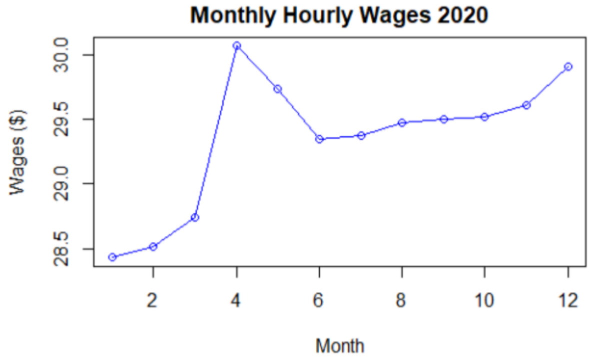
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.85   23.98   25.36   25.71   27.16   29.92 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  23.21   24.29   25.61   25.94   27.32   29.91       1 
```

We can see the summary statistics for the average hourly wages in the months of January and December between 2011 and 2021. Obviously December 2021 would be the one NA from that variable because it has yet to occur. December has a higher minimum and average but January has a higher maximum, probably because the amount came from 2021, whereas December's amount for 2021 has yet to be recorded.

Plot Some Features

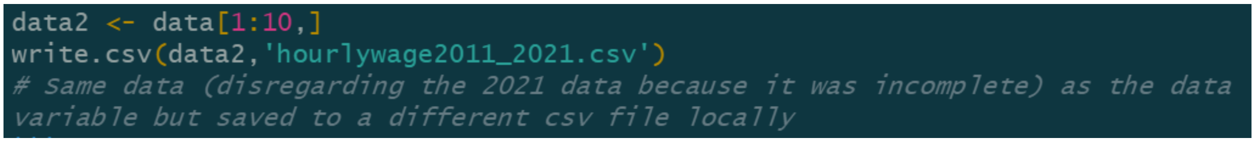






After plotting histograms, I decided they were not the right type of graph to display the data and rather, line plot would better represent because it's showing the data through time. The most interesting find from these graphs would be the dip in 2020 most likely representing the economic result of the Coronavirus. Other than that the graphs showed a pretty steady climb in the average hourly wages.

Save Data Locally as a CSV File



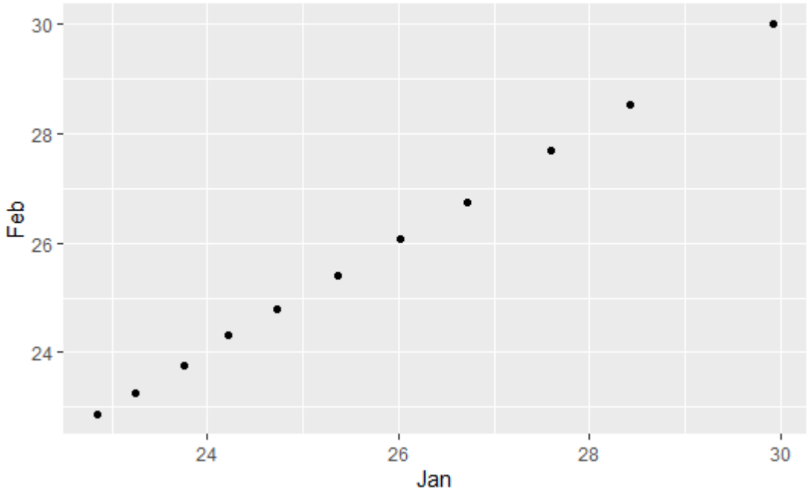
```
data2 <- data[1:10,]
write.csv(data2,'hourlywage2011_2021.csv')
# Same data (disregarding the 2021 data because it was incomplete) as the data
variable but saved to a different csv file locally
```

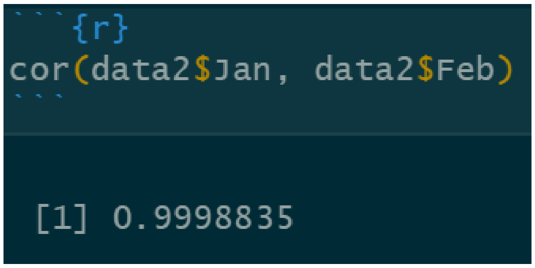
I saved a new file locally without the last row of the data because it did not contain the average hourly rates from June-December 2021. Those NA values would not be included in the new csv file titled hourlywage2011_2020.

Explore Bivariate Relationships

Exploring bivariate relationships between the months does not make a lot of sense data wise because they should be highly correlated with one another as typically the average hourly wage increases every month.

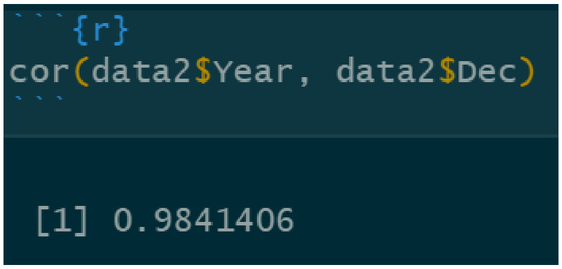
Below is the scatterplot showing the relationship between January and February, which are highly correlated with one another. This is because for both months each year the hourly wage increased.





```{r}
cor(data2$Jan, data2$Feb)
```

```
[1] 0.9998835
```

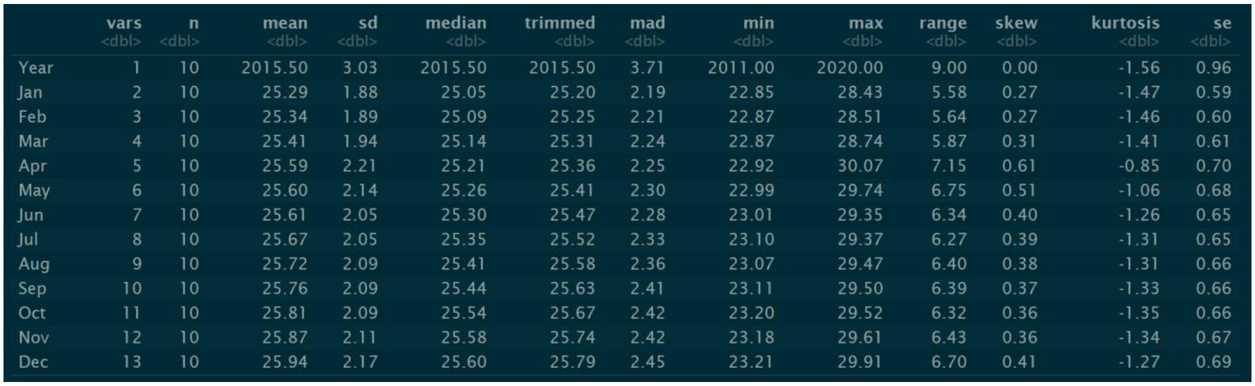


```{r}
cor(data2$Year, data2$Dec)
```

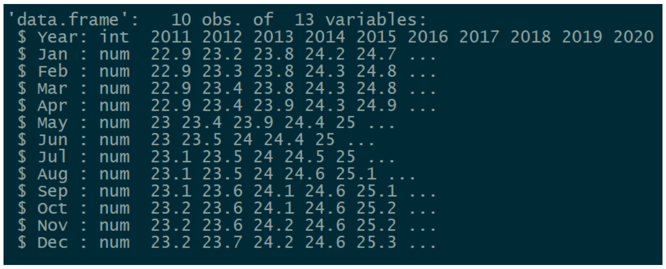
```
[1] 0.9841406
```

These above results print the correlation values between Jan and Feb columns and the Year and Dec columns. As the years increase the average hourly wages in December are also increasing. Again, this is typically a weird bivariate relationship to explore but for the refresher in R I thought the practice would be sufficient.

Summary Report



	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Year	1	10	2015.50	3.03	2015.50	2015.50	3.71	2011.00	2020.00	9.00	0.00	-1.56	0.96
Jan	2	10	25.29	1.88	25.05	25.20	2.19	22.85	28.43	5.58	0.27	-1.47	0.59
Feb	3	10	25.34	1.89	25.09	25.25	2.21	22.87	28.51	5.64	0.27	-1.46	0.60
Mar	4	10	25.41	1.94	25.14	25.31	2.24	22.87	28.74	5.87	0.31	-1.41	0.61
Apr	5	10	25.59	2.21	25.21	25.36	2.25	22.92	30.07	7.15	0.61	-0.85	0.70
May	6	10	25.60	2.14	25.26	25.41	2.30	22.99	29.74	6.75	0.51	-1.06	0.68
Jun	7	10	25.61	2.05	25.30	25.47	2.28	23.01	29.35	6.34	0.40	-1.26	0.65
Jul	8	10	25.67	2.05	25.35	25.52	2.33	23.10	29.37	6.27	0.39	-1.31	0.65
Aug	9	10	25.72	2.09	25.41	25.58	2.36	23.07	29.47	6.40	0.38	-1.31	0.66
Sep	10	10	25.76	2.09	25.44	25.63	2.41	23.11	29.50	6.39	0.37	-1.33	0.66
Oct	11	10	25.81	2.09	25.54	25.67	2.42	23.20	29.52	6.32	0.36	-1.35	0.66
Nov	12	10	25.87	2.11	25.58	25.74	2.42	23.18	29.61	6.43	0.36	-1.34	0.67
Dec	13	10	25.94	2.17	25.60	25.79	2.45	23.21	29.91	6.70	0.41	-1.27	0.69



```
'data.frame':   10 obs. of  13 variables:
 $ Year: int  2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
 $ Jan : num  22.9 23.2 23.8 24.2 24.7 ...
 $ Feb : num  22.9 23.3 23.8 24.3 24.8 ...
 $ Mar : num  22.9 23.4 23.8 24.3 24.8 ...
 $ Apr : num  22.9 23.4 23.9 24.3 24.9 ...
 $ May : num  23 23.4 23.9 24.4 25 ...
 $ Jun : num  23 23.5 24 24.4 25 ...
 $ Jul : num  23.1 23.5 24 24.5 25 ...
 $ Aug : num  23.1 23.5 24 24.6 25.1 ...
 $ Sep : num  23.1 23.6 24.1 24.6 25.1 ...
 $ Oct : num  23.2 23.6 24.1 24.6 25.2 ...
 $ Nov : num  23.2 23.6 24.2 24.6 25.2 ...
 $ Dec : num  23.2 23.7 24.2 24.6 25.3 ...
```

Above we can view summary statistics for every variable in the data set as well as the type of data for each variable. Other than year which is an integer, all of the other columns are numeric type because they are decimals.

This data set was probably not the best to explore for all of these steps, however it was small and simple and the functions could still be applied although not a lot of useful information could be recovered.

I find the increase in average hourly wages to be an interesting topic as to why I chose this one but clearly with the months being the variables and with hourly wages on the rise all variables will be highly correlated.

After taking out the incomplete row for 2021, December had the highest wage value because it was the most recent point on record.

Something that was interesting to view was the range between the variables. Meaning between the years of 2011 and 2020 or ten years worth of time we can see the increase in hourly wages. Choosing one month as an example, December, we can see that from December 2011 to December 2020 the average hourly wage increased by $6.70.

April has the highest standard deviation, meaning it has the most points furthest from it's own mean, while January has the lowest standard deviation. Expectedly April therefore has the highest range and January has the lowest range.