Beer Sales

4/27/2017

Team 26

Dhuri Sayli
Jariwala Ishani
Joshi Aditya
Bhadauria Poornima
Pokale Chetan

Contents

Linear Regression Analysis

We performed linear regression analysis using Weekly Averages Sales Volume of beer as the dependent variable Y and the following as our explanatory variables(X):

1. Age9
2. Age60
3. % College Graduates
4. % With No Vehicles
5. Exponential of Log of Median Income
6. Average Household Size
7. % Working Women with full-time jobs
8. Mean Household Value (Approximated)
9. % of Singles
10. % of Retired
11. % of Unemployed
12. % of working women with children
13. % of non-working women with children
14. % of households with mortgages
15. % of population that is non-white
16. % of population with income under $15,000

All the variables do not have a significant effect on the dependent variable.

Significant Variables

Based on higher co-efficient value (both positive and negative) and lower p-value we have shortlisted the below significant variables

1. Age60
2. % of Unemployed
3. % With No Vehicles
4. % of population with income under $15,000
5. % of working women with children
6. % of Retired
7. % Working Women with full-time jobs

Non-Significant Variables

We have dropped the below variables as we feel they are non-significant because they have lower co-efficient value in our regression analysis and have minimum impact on the dependent variable.

1. % of households with mortgages
2. Mean Household Value (Approximated)
3. % of Singles
4. % of population that is non-white
5. % of non-working women with children
6. % College Graduates
7. Age9
8. Average Household Size

9. Exponential of Log of Median Income

Model Re-Run

We reran the model using five of the below significant variables.

1. Age60
2. % With No Vehicles
3. % of Retired
4. % of Unemployed
5. % of population with income under $15,000

We get an adjusted R square value of 0.03 which is very low and higher p-values as seen below indicating that model might not be linear.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.29848							
R Square	0.08909							
Adjusted R	0.03144							
Standard E	126.323							
Observati	85							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressio	5	123293	24658.6	1.54526	0.18543			
Residual	79	1260651	15957.6					
Total	84	1383944						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	225.025	285.571	0.78798	0.43306	-343.389	793.44	-343.389	793.44
AGE60	-23.5632	1060.61	-0.02222	0.98233	-2134.66	2087.53	-2134.66	2087.53
NOCAR	-565.605	286.887	-1.97153	0.05216	-1136.64	5.42936	-1136.64	5.42936
RETIRED	203.5	1290.77	0.15766	0.87513	-2365.71	2772.71	-2365.71	2772.71
UNEMP	1166.87	1545.11	0.75521	0.45237	-1908.58	4242.33	-1908.58	4242.33
POVERTY	1132.12	1080.74	1.04755	0.29804	-1019.03	3283.27	-1019.03	3283.27

Test of regression assumptions

For performing linear regression there are certain assumptions that should hold true.

Linearity

To perform regression analysis, the relationship between dependent and explanatory variables must be linear which is true in our case. This can be seen by performing Ramsey Regression Equation Specification Error Test (RESET) (1969) to test for linearity:

```
  readXL("//hd.ad.syr.edu/02/258c77/Documents/Downloads/BeerSalesFiltered.xlsx",
   rownames=FALSE, header=TRUE, na="", sheet="BeerSales",
  stringsAsFactors=TRUE)
RegModel.3 <- lm(WEEKVOL~AGE60+NOCAR+POVERTY+RETIRED+UNEMP, data=Dataset)
summary(RegModel.3)

resettest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP, power=2:3,
  type="regressor", data=Dataset)
```

Submit

Output

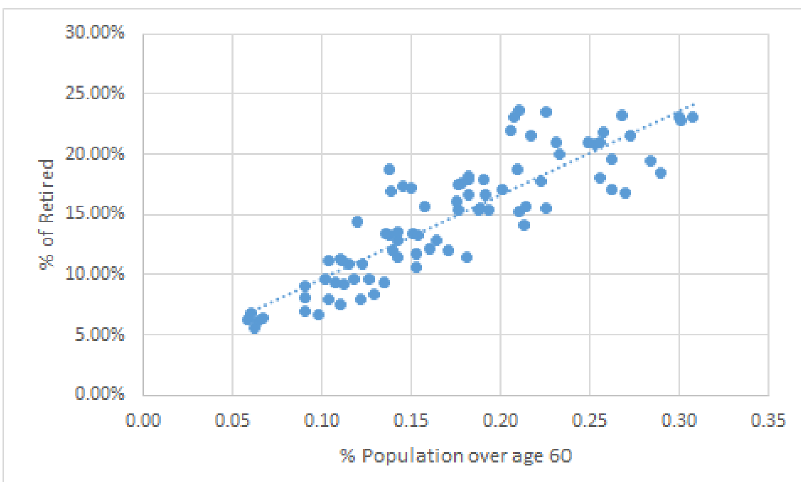
```
RETIRED       203.50    1290.77   0.158    0.8751
UNEMP        1166.87    1545.11   0.755    0.4524
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 126.3 on 79 degrees of freedom
Multiple R-squared:  0.08909, Adjusted R-squared:  0.03144
F-statistic: 1.545 on 5 and 79 DF,  p-value: 0.1854


> resettest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP, power=2:3,
+   type="regressor", data=Dataset)

        RESET test

data:  WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP
RESET = 0.9111, df1 = 10, df2 = 69, p-value = 0.5281
```

From the above tests it can be seen that p-value is 0.5281 which is greater than 0.05 and there is no linearity problem. So this assumption is not violated.

Correlation of X variables

X variables or the explanatory variables should not be correlated with each other. However there is a relationship between some of our X variables and multi-collinearity exists as seen from the below plots and Variance Inflation Factor test of correlated explanatory variables tests.

	WEEKVOL	*AGE60*	*NOCAR*	*RETIRED*	*UNEMP*	*POVERTY*
WEEKVOL	1					
AGE60	-0.05586	1				
NOCAR	-0.05655	0.178627	1			
RETIRED	0.024966	0.871631	0.472909	1		
UNEMP	0.167446	-0.36863	0.560859	0.084464	1	
POVERTY	0.051973	0.162991	0.917598	0.527953	0.686419	1



```
    rownames=FALSE, header=TRUE, na="", sheet="BeerSales",
  stringsAsFactors=TRUE)
RegModel.3 <- lm(WEEKVOL~AGE60+NOCAR+POVERTY+RETIRED+UNEMP, data=Dataset)
summary(RegModel.3)

resettest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP, power=2:3,
  type="regressor", data=Dataset)
vif(RegModel.2)
```

Output

Submit

```
> resettest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP, power=2:3,
+   type="regressor", data=Dataset)

        RESET test

data:  WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP
RESET = 0.9111, df1 = 10, df2 = 69, p-value = 0.5281


> vif(RegModel.3)
    AGE60     NOCAR   POVERTY   RETIRED     UNEMP
23.772027  7.023441 11.643476 22.471997  6.552314

> vif(RegModel.2)
    AGE60     NOCAR   POVERTY   RETIRED     UNEMP
23.772027  7.023441 11.643476 22.471997  6.552314
```

From the above tests it can be seen that variation influence factors are more than 10 for three variables Age60, Poverty and Retired which proves that multi collinearity exists and this assumption is violated.

Heteroscedasticity

The error term must have constant variance over a range of X values. In our case the size of error term does not depend on any explanatory variable and there is no heteroscedasticity as seen from the below Breusch-Pagan test below:

```
vif(RegModel.2)


RegModel.4 <- lm(WEEKVOL~AGE60+NOCAR+POVERTY+RETIRED+UNEMP, data=Beer)
summary(RegModel.4)

bptest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP, varformula = ~
  fitted.values(RegModel.4), studentize=FALSE, data=Beer)
```

Output

Submit

```
RETIRED        203.50    1290.77    0.158    0.8751
UNEMP         1166.87    1545.11    0.755    0.4524
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 126.3 on 79 degrees of freedom
Multiple R-squared:  0.08909, Adjusted R-squared:  0.03144
F-statistic: 1.545 on 5 and 79 DF,  p-value: 0.1854


> bptest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP, varformula = ~
+   fitted.values(RegModel.4), studentize=FALSE, data=Beer)

        Breusch-Pagan test

data:  WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP
BP = 0.022753, df = 1, p-value = 0.8801
```

As the p-value is 0.8801 which is more than 0.05, there is no problem with heteroscedasticity and the assumption is not violated.

Error terms correlation

Error terms must not be correlated over the time period. For our regression analysis there is a problem with serial correlation as seen from the Durbin-Watson test below because p-value is 9.462e-06 which is less than 0.05:

```
RegModel.4 <- lm(WEEKVOL~AGE60+NOCAR+POVERTY+RETIRED+UNEMP, data=Beer)
summary(RegModel.4)

bptest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP, varformula = ~
  fitted.values(RegModel.4), studentize=FALSE, data=Beer)
dwtest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP,
  alternative="greater", data=Beer)
```

Submit

Output

```
> bptest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP, varformula = ~
+   fitted.values(RegModel.4), studentize=FALSE, data=Beer)

        Breusch-Pagan test

data:  WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP
BP = 0.022753, df = 1, p-value = 0.8801


> dwtest(WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP,
+   alternative="greater", data=Beer)

        Durbin-Watson test

data:  WEEKVOL ~ AGE60 + NOCAR + POVERTY + RETIRED + UNEMP
DW = 1.1214, p-value = 9.462e-06
alternative hypothesis: true autocorrelation is greater than 0
```

Hence this assumption is violated.

Outliers

Ideally there should not be any outliers for performing linear regression. However outliers exists in our model which can be seen in the linear scatter plots above and can be proved from the Bonferroni outlier test below.

```
Beer <- readXL("//hd.ad.syr.edu/02/258c77/Documents/Desktop/Beer.xlsx",
  rownames=FALSE, header=TRUE, na="", sheet="Final Data",
  stringsAsFactors=FALSE)
RegModel.2 <- lm(WEEKVOL~AGE60+NOCAR+POVERTY+RETIRED+UNEMP, data=Beer)
summary(RegModel.2)

outlierTest(RegModel.2)
```

Submit

Output

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   225.03     285.57   0.788   0.4331
AGE60         -23.56    1060.61  -0.022   0.9823
NOCAR        -565.60     286.89  -1.972   0.0522 .
POVERTY      1132.12    1080.74   1.048   0.2980
RETIRED       203.50    1290.77   0.158   0.8751
UNEMP        1166.87    1545.11   0.755   0.4524
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 126.3 on 79 degrees of freedom
Multiple R-squared:  0.08909,	Adjusted R-squared:  0.03144
F-statistic: 1.545 on 5 and 79 DF,  p-value: 0.1854

> outlierTest(RegModel.2)
   rstudent unadjusted p-value Bonferonni p
75 3.601553         0.00055434     0.047119
```

It can be seen that there is one outlier which is row number 75 and this assumption is violated.

Regression after correction

Some corrections have been made for the assumptions that were violated.

Linearity

As there was no linearity problem and the assumption was not violated Box-Cox was not performed and there is no need for transformation of Y.

Multicollinearity

As multicollinearity existed between three variables Age60, Poverty and Retired, average was taken for these three and was converted into single variable named AGE_POV_RET as seen in the Rerun Regression tab in the excel.

Heteroscedasticity

Assumption was not violated and hence no correction was performed.

Serial correlation

Though the assumption was violated, it is out of scope for this assignment.

Outliers

One outlier was detected as per Bonferroni outlier test which was the record with weekly sales volume as 875. That particular record has been dropped.

After making the above corrections, linear regression was rerun and the following output was obtained.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.308316
R Square	0.095059
Adjusted R Squ	0.061124
Standard Error	117.1715
Observations	84

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	115373.5	38457.85	2.801181	0.045190145
Residual	80	1098332	13729.15		
Total	83	1213706			

	Coefficients	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	56.38573	142.9986	0.39431	0.694402	-228.1905648	340.962027	-228.19056	340.962027
NOCAR	-370.108	157.4385	-2.35081	0.021195	-683.4203681	-56.795234	-683.42037	-56.795234
UNEMP	1930.411	722.4159	2.672159	0.00913	492.7570112	3368.06401	492.757011	3368.06401
AGE_POV_RET	746.0732	385.7364	1.934153	0.05663	-21.56670408	1513.71318	-21.566704	1513.71318

Adjusted R square improved slightly but is still on the lower side because of the small size of the data and real world fluctuations.

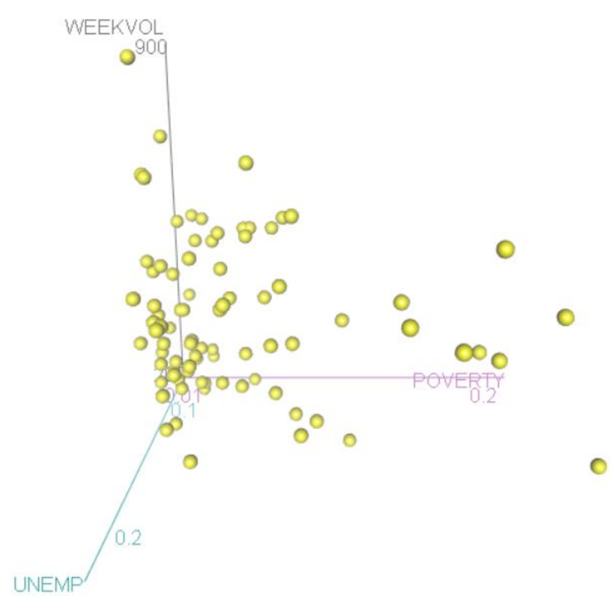
3D Graph using R

```
load("C:/Users/Poornima Bhadauria/Desktop/BeerSalesR.xls")
BeerSales <- readXL("C:/Users/Poornima Bhadauria/Desktop/BeerSalesR.xls",
  rownames=FALSE, header=TRUE, na="", sheet="Sheet1", stringsAsFactors=FALSE)
library(rgl, pos=14)
library(nlme, pos=15)
library(mgcv, pos=15)
scatter3d(WEEKVOL~POVERTY+UNEMP, data=BeerSales, surface=FALSE,
  residuals=TRUE, bg="white", axis.scales=TRUE, grid=TRUE, ellipsoid=FALSE)
```

Submit

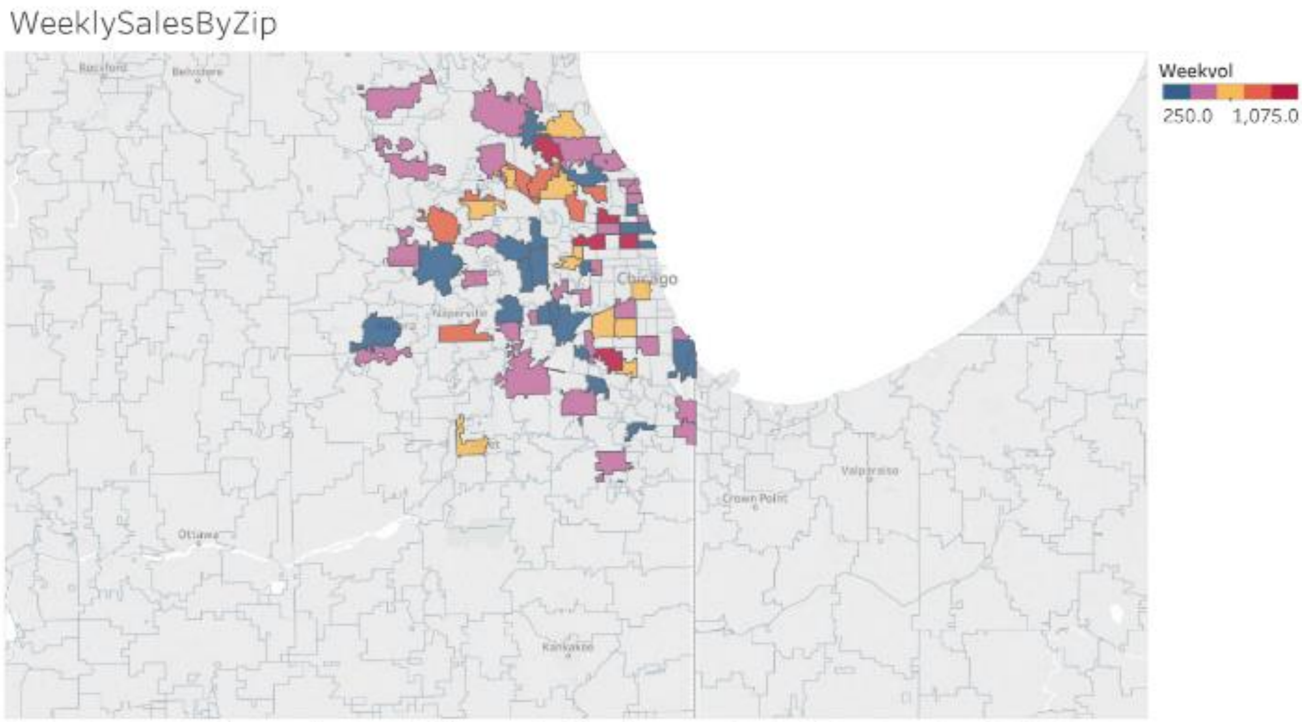
Output

```
> load("C:/Users/Poornima Bhadauria/Desktop/BeerSalesR.xls")

> BeerSales <- readXL("C:/Users/Poornima Bhadauria/Desktop/BeerSalesR.xls",
+   rownames=FALSE, header=TRUE, na="", sheet="Sheet1", stringsAsFactors=FALSE)

> library(rgl, pos=14)

> library(nlme, pos=15)

> library(mgcv, pos=15)

> scatter3d(WEEKVOL~POVERTY+UNEMP, data=BeerSales, surface=FALSE,
+   residuals=TRUE, bg="white", axis.scales=TRUE, grid=TRUE, ellipsoid=FALSE)
```



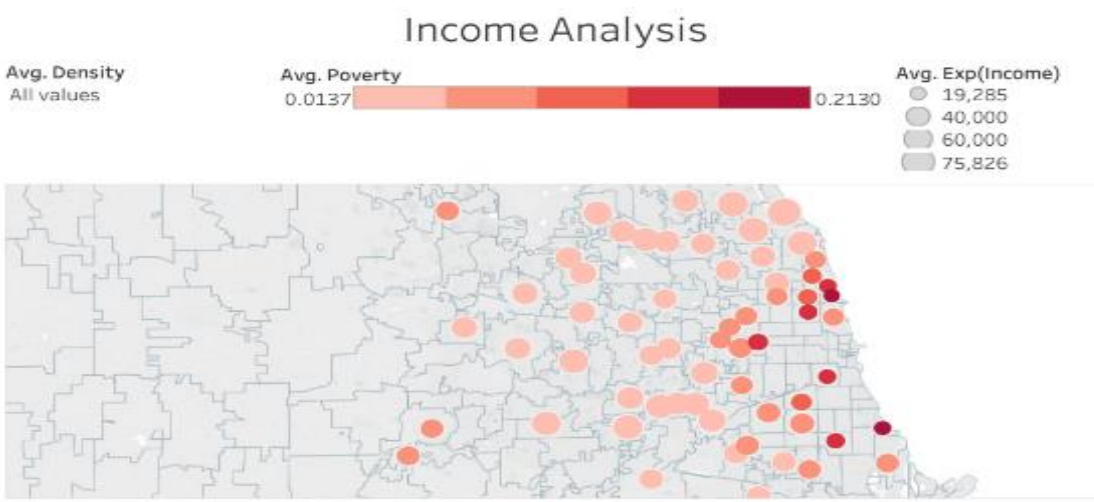
Geographic Representations

Plot 1: This plot describes weekly average sales volume by location. The highest sales is highlighted by red and the lowest sales is highlighted by blue as seen in the map



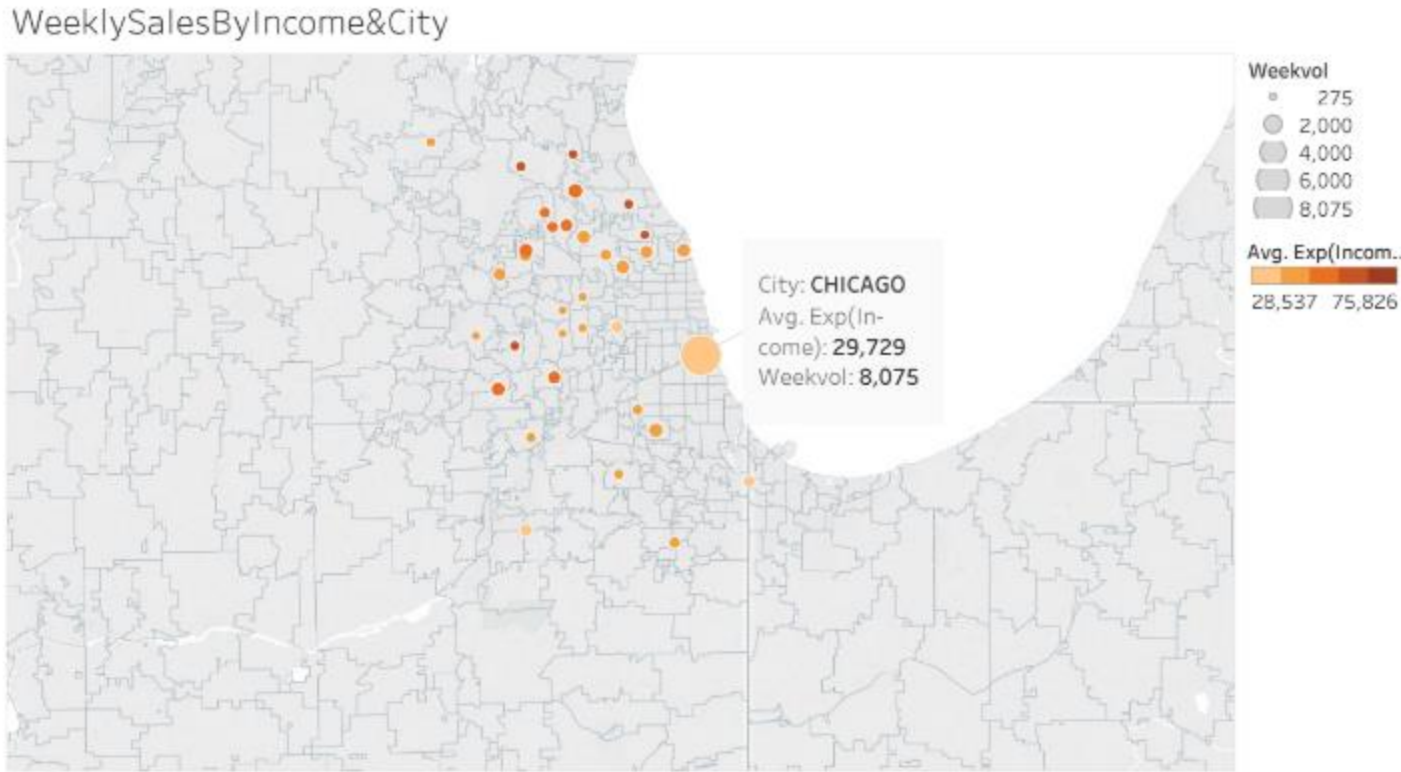
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Weekvol. Details are shown for ZIP. The data is filtered on City, which keeps 59 of 59 members.

Plot 2: This plot tries to analyze how Income, Density and Poverty affects Weekly sales of stores. As seen the region "Aurora" has higher average Poverty and lower average income, hence the sales is low in that region.



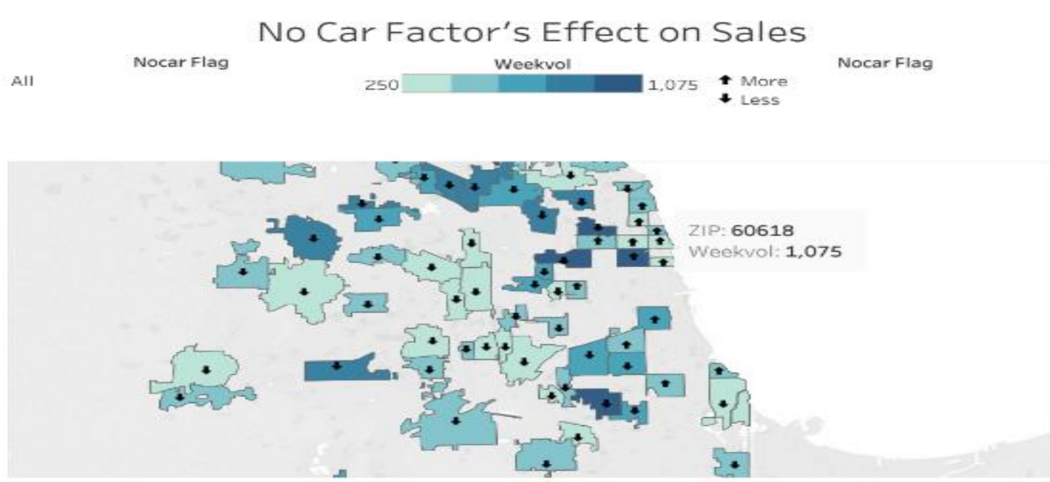
City	Avg. Density	Avg. Poverty	Exp(Income)	Weekly Sales Volume
ADDISON	0.15%	3.26%	41,578.000..	350
ARLINGTON HEIGHTS	0.06%	2.39%	54,694.000..	750
AURORA	0.28%	6.35%	35,730.000..	400
BANNOCKBURN	0.19%	2.32%	75,598.999..	700
BENSENVILLE	0.13%	3.75%	38,170.000..	375
BLOOMINGDALE	0.14%	3.32%	46,374.999..	475
BOLINGBROOK	0.18%	3.31%	39,547	500

Plot 3: This plot shows weekly sales by Income and city. As highlighted, Chicago has highest sales with low Income.



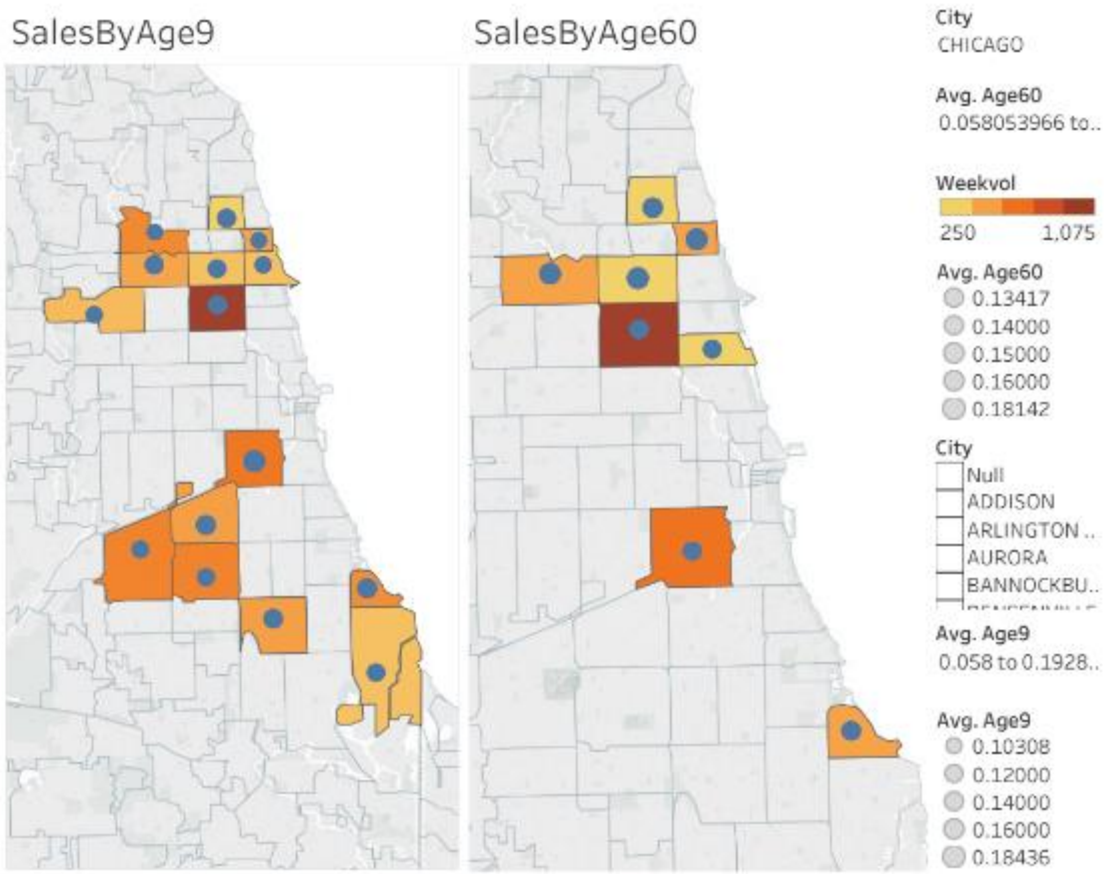
Map based on Longitude (generated) and Latitude (generated). Color shows average of Exp(Income). Size shows sum of Weekvol. Details are shown for City. The view is filtered on City, which keeps 59 of 59 members.

Plot 4: This plot shows how No Car factor affects the sales per each store in particular region. As highlighted, River Forest has less sales, as there are only 12.46% people who don't have car. At the same time, Park Ridge has almost the same sales as River Forest, but it has low percentage of people having no car. Probably the location of store Dominick's 4 is located in the highly populated region and the store is within the walking distance.



Name	City	Nocar	Weekly Sales Volume
DOMINICKS 2	RIVER FOREST	12.46%	350
DOMINICKS 4	PARK RIDGE	5.56%	300
DOMINICKS 5	PALATINE	2.56%	550
DOMINICKS 8	OAK LAWN	7.51%	600

Plot 5: This plot shows the comparison of beer bought by Age groups 9 and 60.



Business Analysis

Looking at analytical patterns from a business perspective, we arrive at the following conclusions

- <u>People with vehicles lead to a higher volume of sales</u>
  People with vehicles have an easy and convenient access to the beer stores thereby increasing the sales.

- <u>People with income lower than 15000USD tend to make a bulk of the sales</u>
  This might indicate the fact people with higher income might go for costly liquor like Whiskey, Scotch or Vodka and people with lower income might be opting for beer as it is comparatively cheaper.

- <u>Sales increase slightly with an increase in the amount of retired customer</u>
  Retired customers have more leisure and free time than people who are working. This might be the reason that they are more inclined towards purchasing beer.

- <u>Sales tend to rise with the increase in the number of unemployed customers</u>
  Unemployed customers have both free time and monetary issues. So beer seems to be the most affordable option.

- <u>Sales tend to increase when the percentage of population below 60 years is less.</u>
  If the percentage op population below 60 years is less it is highly likely that people are retired and have more free time.

Other Factors

1. **Percentage of Unemployed Males**: Unemployed males tend to be a dedicated client base for beer purchases. Having this data would allow us to find ways and run campaigns to better target them.
2. **Holiday Sales Data**: Specific sales data for various holidays, such as Christmas and new years, etc would allow us to figure out which seasonal campaign was successful and when improvement is needed.
3. **Promotional event sales**: Sales data for any promotional campaign usually results in spikes of sales. Having this data would allow us to better realize the marketability and hence sales prediction of the beer.
4. **Unfulfilled request count:** Amount of times a customer was unable to find the beer of his choice would enable us to keep a smarter stock and help us make better predictions
5. **Sales timings**: Knowing what time attracts the most customers and comparing them with regular store timings can help us predict how the sales are going to be.

Excel File

BeerSalesFinal.xlsx