Initial EDA

Mike Finnegan

9/217/2017

ACS Median Household Income Data for Birmingham-Hoover MSA

After first importing data in from an excel file holding data taken from the ACS, I create a few graphs showing how median household income varies by race and age in the Birmingham-Hoover MSA.

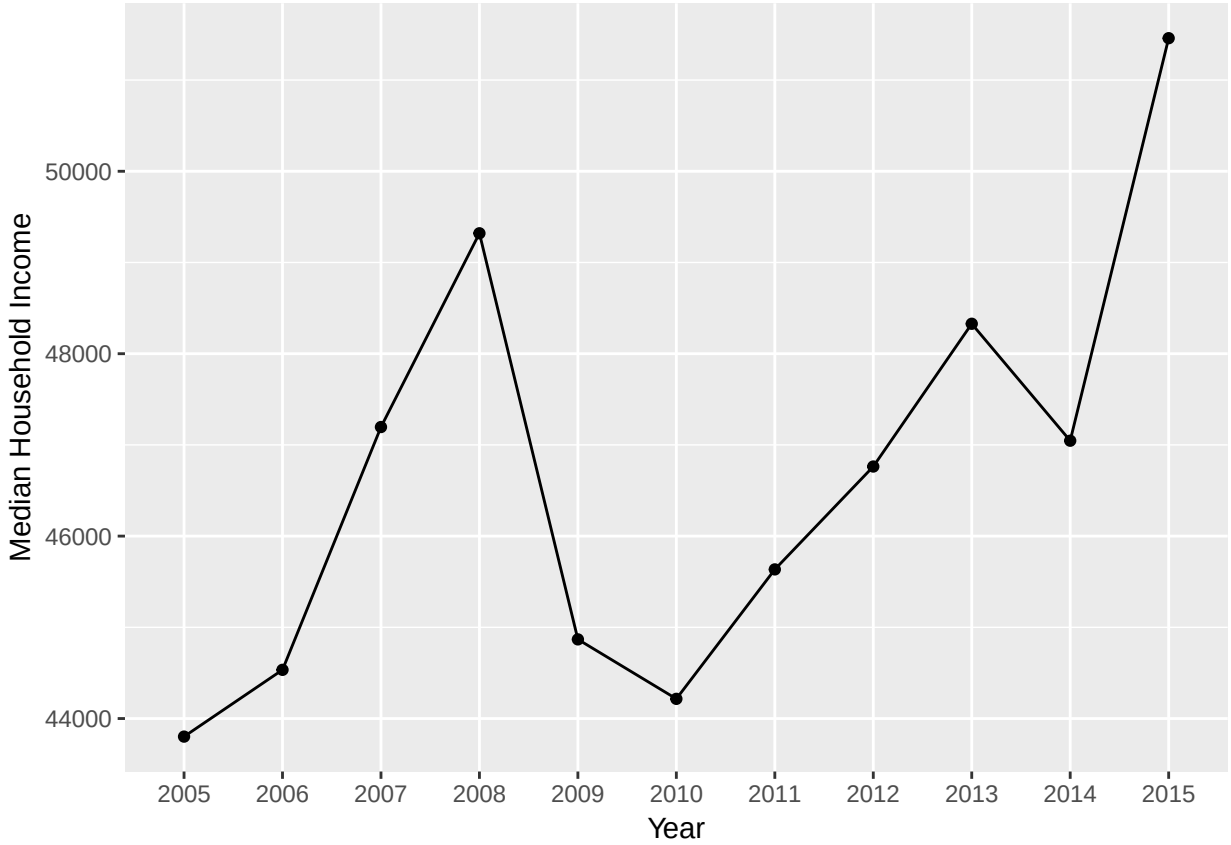
Import Data extracted from ACS via Excel

```
library(readxl)
Bham_Median_HH_Income <- read_excel(
"~/Documents/Github/BirminghamGentrification/Bham_Median_HH_Income.xlsx")
Bham_MarginofError <- read_excel(
"~/Documents/Github/BirminghamGentrification/Bham_Median_HH_Income_MarginofError.xlsx")
Bham_Median_HH_Income$Year <- as.factor(Bham_Median_HH_Income$Year)
```

Plot Median Household Incomes from 2005 to 2015

```
library(ggplot2)

#Plot line chart Overall Household Income for Birmingham-Hoover MSA
ggplot(Bham_Median_HH_Income, aes(x=Year,`Overall Households`, group=1)) +
  geom_line() +
  geom_point() +
  xlab("Year") +
  ylab("Median Household Income")
```



The 2008 Recession stunts what appears to be significant growth occuring betwen years 2005 and 2008, essentially putting the Birmingham area's median household income level back 5 years (nota bene: compare 2010 and 2005). However, post-recession median household income has risen steadily with the largest growth in the most recent year for which data is available. Intuition from living in the area tells us that this rate of growth may have continued to increase in the past year and a half.

Race and Age

```
#Plot Median Household Income by Race
ggplot(Bham_Median_HH_Income, aes(x=Year, group=1)) +
  geom_line(aes(y=`Overall Households`, colour="Overall")) +
  geom_line(aes(y=`Race: white`, colour="White")) +
  geom_line(aes(y=`Race: black`, colour="Black")) +
  geom_line(aes(y=`Race: hispanic`, colour="Hispanic")) +
  scale_colour_manual("",
    values = c("Overall"="black", "White"="green",
    "Black"="blue", "Hispanic"="purple")) +
  xlab("Year") +
  ylab("Median Household Income")
```

Looking at the graph we can see that in the 2005 to 2010 time from the median income of white households has risen by roughly $10k, while the median income of black households has risen by approximately $5k. However, the two groups' levels of income have moved in roughly the same manner. Meanwhile, the median income for Hispanic households is largely volatile and non-cyclical in relation to broader economic treads.

```
#Plot Median Household Income for Age
ggplot(Bham_Median_HH_Income, aes(x=Year, group=1)) +
  geom_line(aes(y=`Overall Households`, colour="Overall")) +
```

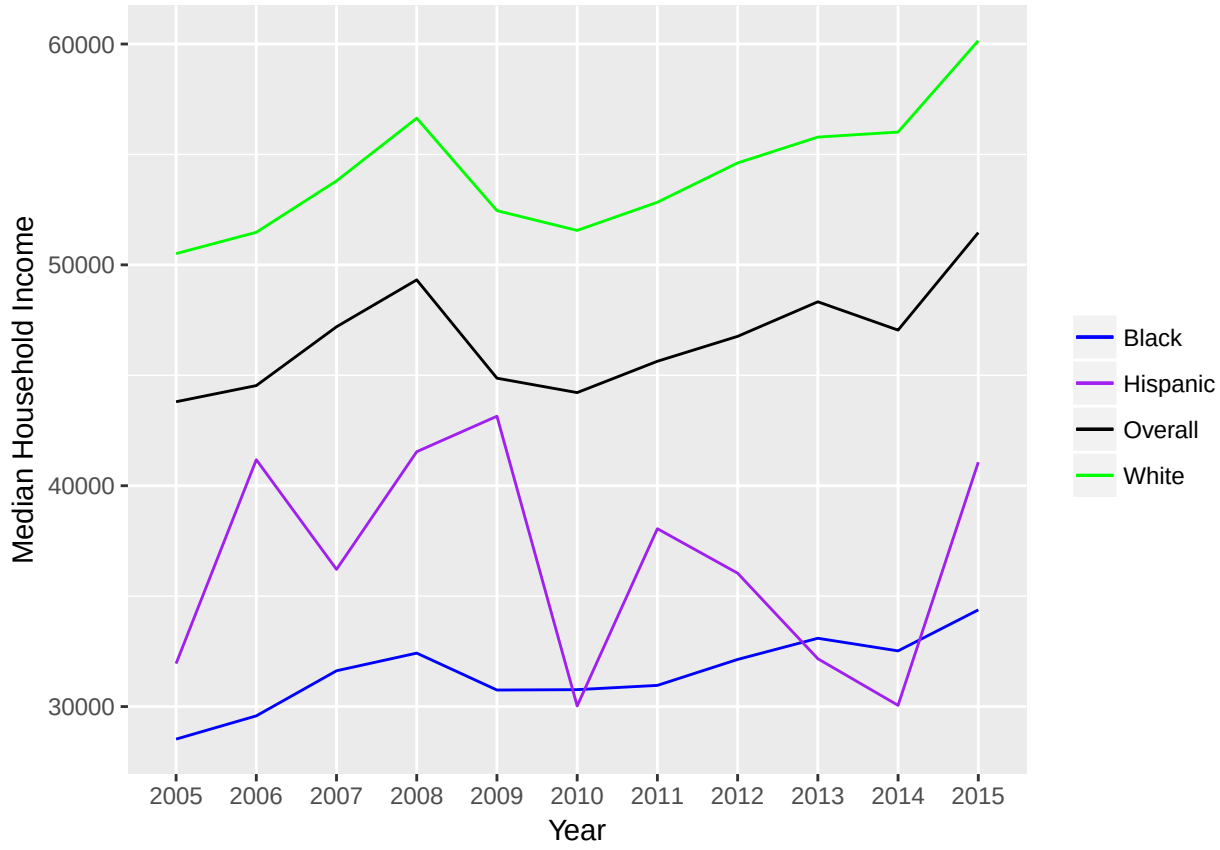


Figure 1: Median Household Income by Race.

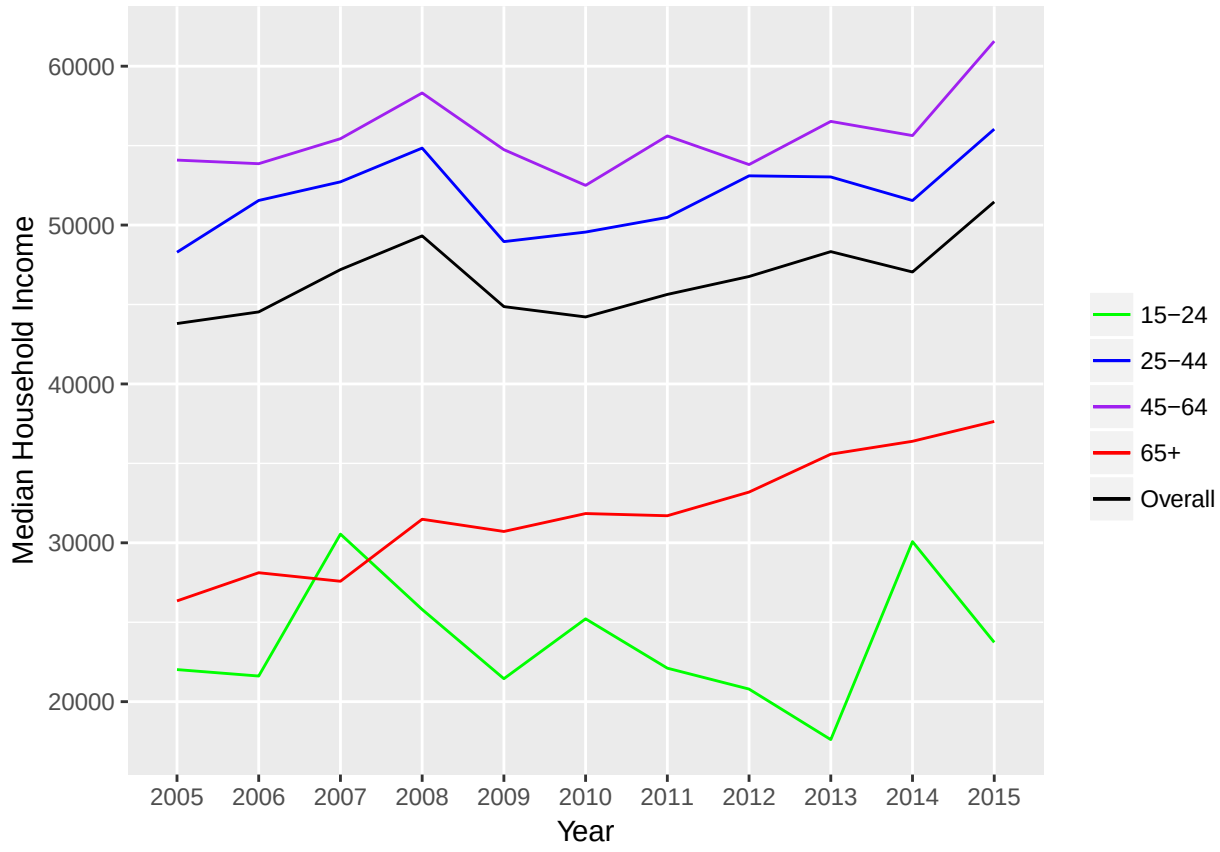


Figure 2: Median Houshold Income by Age.

```
  geom_line(aes(y=`Age: 15 to 24 years`, colour="15-24")) +
  geom_line(aes(y=`Age: 25 to 44 years`, colour="25-44")) +
  geom_line(aes(y=`Age: 45 to 64 years`, colour="45-64")) +
  geom_line(aes(y=`Age: 65 years and over`, colour="65+")) +
  scale_colour_manual("",
    values = c("Overall"="black", "15-24"="green",
      "25-44"="blue", "45-64"="purple",
      "65+"="red")) +
  xlab("Year") +
  ylab("Median Household Income")
```

We would expect to see an increase in median household incomes for age groups 15-24 and 25-44 (more so for the latter), but there is little evidence of this seen in this graph.

Map Median Household Income in Jefferson County, Alabama

In this section I begin visualizing the changes in Median Household Income within the Birmingham metropolitan area (specifically Jefferson County). By utilzing the tidycensus package, I use an api to directly access ACS data from the 2010 and 2015 surveys. From there I use leaflet amongst other packages to create a map of Jefferson county. The 2010 and 2015 maps are displayed, but the focus of this section is the percent change in median household income, i.e. the last map shown.

Install necessary packages

```
#Begin work with tidycensus
library("tidycensus")
library("tidyverse")
library("leaflet")
library("sf")
library("stringr")
library("viridis")
library("viridisLite")
library("acs")
```

Link api key

```
#census_api_key("c2d75e7e54dd23544e0d77f8d8b98819f00ccbb3", install=TRUE)
readRenviron("~/.Renviron")
```

Jefferson County in 2010

The code in order to produce the map is shown here for the reader's interest and for the sake of reproducibility. Due to length and redundancy reasons, the code to create maps will not be included for the ones following this. The code begins by using the get_acs function to pull and create the data set we want. From there I transform the coordinates of each census tract included in the ACS dataset to a workable format and then create a legend box showing the values associated with the color scale.

```
#load variables table to find relevant ACS table
#v15 <- load_variables(2015, "acs5", cache = TRUE)
#View(v15)

Jeff_HH_income10 <- get_acs(geography = "tract",
                            variables = "B19013_001E",
                            endyear=2010,
                            state = "AL",
                            county = "Jefferson County",
                            geometry = TRUE)

pal <- colorNumeric(palette = "viridis",
                    domain = Jeff_HH_income10$estimate)

Jeff_HH_income10 %>%
  st_transform(crs = "+init=epsg:4326") %>%
  leaflet(width = "100%") %>%
  addProviderTiles(provider = "CartoDB.Positron") %>%
  addPolygons(popup = ~ str_extract(estimate, "^([^,]*)"),
              stroke = FALSE,
              smoothFactor = 0,
              fillOpacity = 0.7,
              color = ~ pal(estimate)) %>%
  addLegend("bottomright",
            pal = pal,
            values = ~ estimate,
            title = "2010 Median Household Income",
            labFormat = labelFormat(prefix = "$"),
            opacity = 1)
```

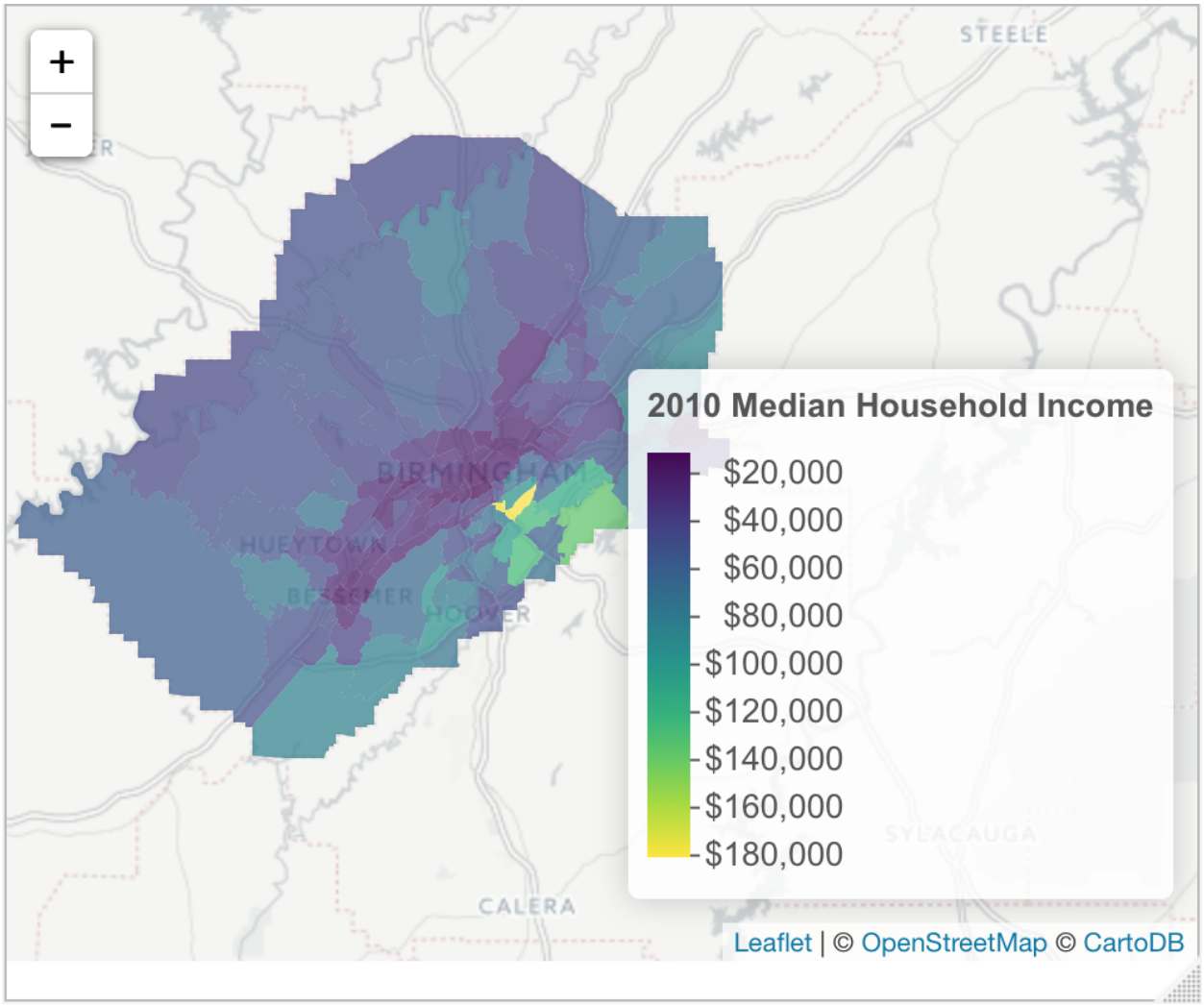


Figure 3:

2010 Jefferson County Median Household Income by Census Tract.

Jefferson County in 2015

2015 Jefferson County Median Household Income by Census Tract.

Prepare Data for Percent Change in Median Household Income

```
#sort tables for computation
Jeff_HH_income10 <- Jeff_HH_income10[order(Jeff_HH_income10$GEOID),]
Jeff_HH_income15 <- Jeff_HH_income15[order(Jeff_HH_income15$GEOID),]

#compute percentage change
Jeff_HH_income15$estimate2010 <- Jeff_HH_income10$estimate
Jeff_HH_income15$percent_change <- ((Jeff_HH_income15$estimate -
                    Jeff_HH_income15$estimate2010)/Jeff_HH_income15$estimate2010)*100
```

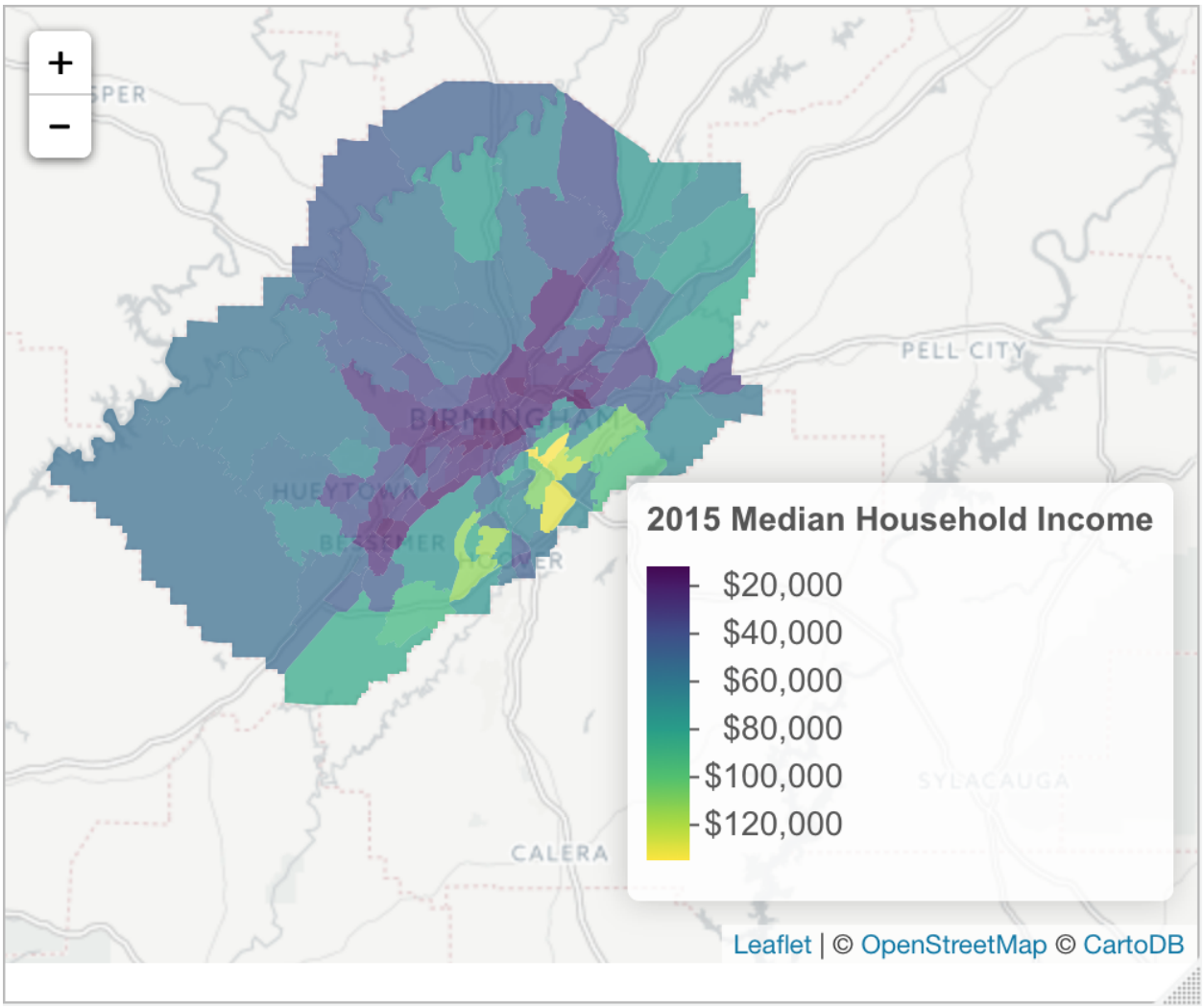

+
−
PELL CITY
BIRMINGHAM
HUEYTOWN
SYLACAUGA
CALERA
2015 Median Household Income
$20,000
$40,000
$60,000
$80,000
$100,000
$120,000
Leaflet | © OpenStreetMap © CartoDB

Figure 4:

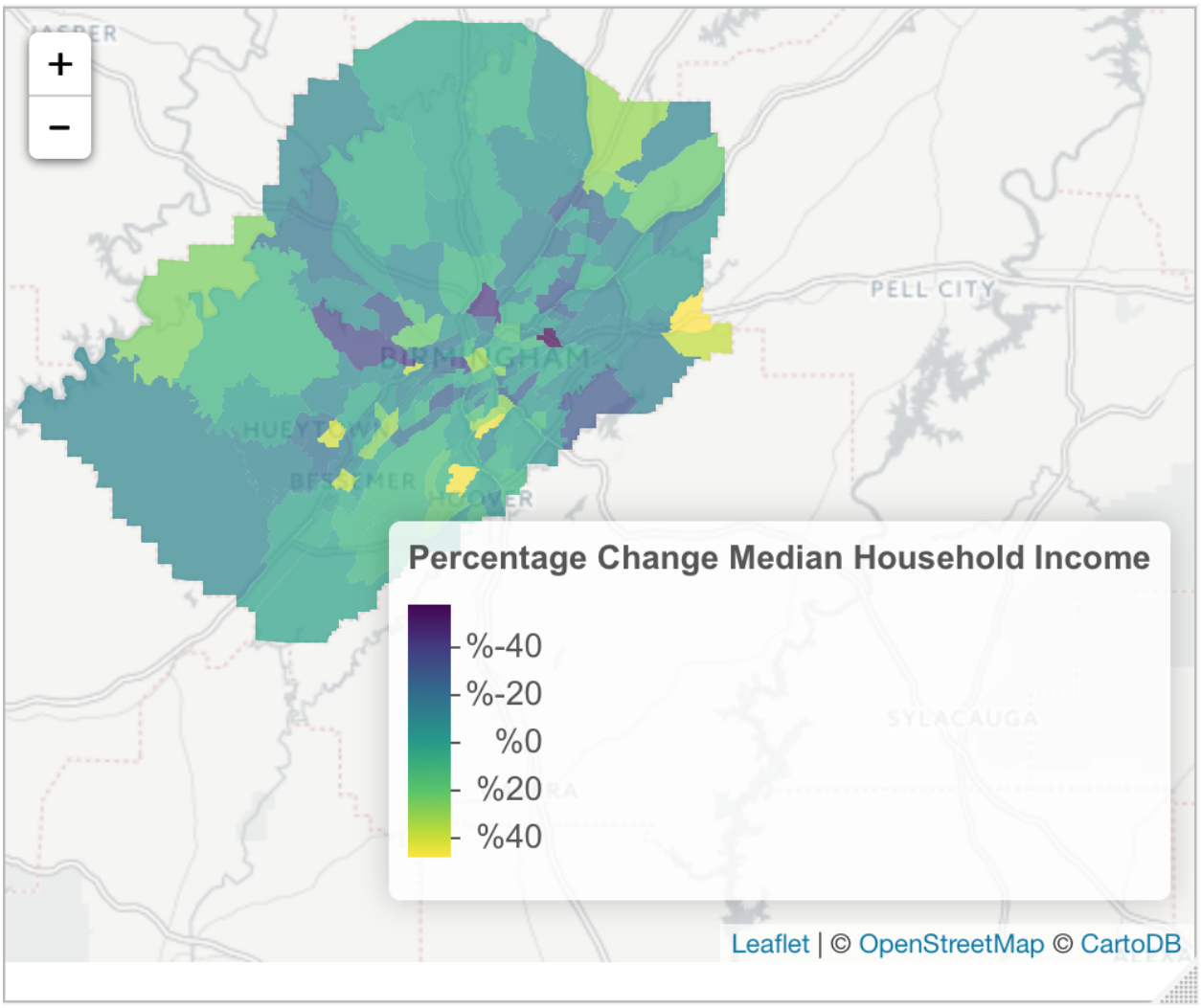


Figure 5:

```
#check to make sure geometry column is of object class "sf" in order for
#st_transform to be able to map coordinates to map
class(Jeff_HH_income15)
#checks out okay
```

Percent Change Map

Percentage Change in Median Household Income from 2010 to 2015 by Census Tract.

The map above detailing the percent change in median household income from 2010 to 2015 provides some insight. As one would expect, growing neighbordhoods like Avondale, Highland park,and 1st Ave North have positive percentage change. However, each of these neighborhoods has an adjacent neighborhood that has experienced a strong negetive percentage change. These neighborhoods are primarily Oak Ridge Park, Graymont, and Irondale.

One possible explanation for this phenomenon could be that resources flowing into the growing census tracts are disproportionately coming from certain adjacent census tracts, which are those that are seen having experienced large negative declines in median household income. Another perspective of this same explanation

is that the low-income households being forced out of the growing neigborhoods are concentrating themselves in the declining household income census tracts.

Further questions from this are: 1) If low-income households are concentrating in particular neighborhoods, what factors are driving this concentration? Following this, what are the characteristics of these neighborhoods 2) Similarly, what are the characteristics of the growing neighborhoods? What is the marginal effect of the presence of a brewery? Is there a necessary time period the brewery must be open before other business open around it?

QCEW analysis

```
library(blsAPI)

Q4Y16 <- blsQCEW('Area', year='2016', quarter='4', area='01073')
#NAICS codes:
#all- 10 & own_code=5
#breweries- 31212
#restaurants and bars-722 & own_code=5
```

Pull QCEW data

```
#For Q4Y16i
Q4Y16 <- blsQCEW('Area', year='2016', quarter='4', area='01073')
Q4Y16i <- rbind(
  subset(Q4Y16, industry_code==10 & own_code==5),
  subset(Q4Y16, industry_code==31212),
  subset(Q4Y16, industry_code==722 & own_code==5),
  subset(Q4Y16, industry_code==72233 & own_code==5))

#For Q3Y16i
Q3Y16 <- blsQCEW('Area', year='2016', quarter='3', area='01073')
Q3Y16i <- rbind(
            subset(Q3Y16, industry_code==10 & own_code==5),
            subset(Q3Y16, industry_code==31212),
            subset(Q3Y16, industry_code==722 & own_code==5),
            subset(Q3Y16, industry_code==72233 & own_code==5))

#For Q2Y16i
Q2Y16 <- blsQCEW('Area', year='2016', quarter='2', area='01073')
Q2Y16i <- rbind(
  subset(Q2Y16, industry_code==10 & own_code==5),
  subset(Q2Y16, industry_code==31212),
  subset(Q2Y16, industry_code==722 & own_code==5),
  subset(Q2Y16, industry_code==72233 & own_code==5))

Q4Y16 <- blsQCEW('Area', year='2016', quarter='4', area='01073')
Q3Y16 <- blsQCEW('Area', year='2016', quarter='3', area='01073')
Q2Y16 <- blsQCEW('Area', year='2016', quarter='2', area='01073')
Q1Y16 <- blsQCEW('Area', year='2016', quarter='1', area='01073')

#For Q1Y16i
Q1Y16i <- rbind(
        subset(Q1Y16, industry_code==10 & own_code==5),
```

```
        subset(Q1Y16, industry_code==31212),
        subset(Q1Y16, industry_code==722 & own_code==5),
        subset(Q1Y16, industry_code==72233 & own_code==5))
y16i <- rbind(Q4Y16i,Q3Y16i,Q2Y16i,Q1Y16i)

Q4Y15 <- blsQCEW('Area', year='2015', quarter='4', area='01073')
Q3Y15 <- blsQCEW('Area', year='2015', quarter='3', area='01073')
Q2Y15 <- blsQCEW('Area', year='2015', quarter='2', area='01073')
Q1Y15 <- blsQCEW('Area', year='2015', quarter='1', area='01073')

Q4Y15i <- rbind(
  subset(Q4Y15, industry_code==10 & own_code==5),
  subset(Q4Y15, industry_code==31212),
  subset(Q4Y15, industry_code==722 & own_code==5),
  subset(Q4Y15, industry_code==72233 & own_code==5))
Q3Y15i <- rbind(
  subset(Q3Y15, industry_code==10 & own_code==5),
  subset(Q3Y15, industry_code==31212),
  subset(Q3Y15, industry_code==722 & own_code==5),
  subset(Q3Y15, industry_code==72233 & own_code==5))
Q2Y15i <- rbind(
  subset(Q2Y15, industry_code==10 & own_code==5),
  subset(Q2Y15, industry_code==31212),
  subset(Q2Y15, industry_code==722 & own_code==5),
  subset(Q2Y15, industry_code==72233 & own_code==5))
Q1Y15i <- rbind(
  subset(Q1Y15, industry_code==10 & own_code==5),
  subset(Q1Y15, industry_code==31212),
  subset(Q1Y15, industry_code==722 & own_code==5),
  subset(Q1Y15, industry_code==72233 & own_code==5))
y15i <- rbind(Q4Y15i,Q3Y15i,Q2Y15i,Q1Y15i)



Q4Y14 <- blsQCEW('Area', year='2014', quarter='4', area='01073')
Q3Y14 <- blsQCEW('Area', year='2014', quarter='3', area='01073')
Q2Y14 <- blsQCEW('Area', year='2014', quarter='2', area='01073')
Q1Y14 <- blsQCEW('Area', year='2014', quarter='1', area='01073')

Q4Y14i <- rbind(
  subset(Q4Y14, industry_code==10 & own_code==5),
  subset(Q4Y14, industry_code==31212),
  subset(Q4Y14, industry_code==722 & own_code==5),
  subset(Q4Y14, industry_code==72233 & own_code==5))
Q3Y14i <- rbind(
  subset(Q3Y14, industry_code==10 & own_code==5),
  subset(Q3Y14, industry_code==31212),
  subset(Q3Y14, industry_code==722 & own_code==5),
  subset(Q3Y14, industry_code==72233 & own_code==5))
Q2Y14i <- rbind(
  subset(Q2Y14, industry_code==10 & own_code==5),
  subset(Q2Y14, industry_code==31212),
  subset(Q2Y14, industry_code==722 & own_code==5),
```

```
  subset(Q2Y14, industry_code==72233 & own_code==5))
Q1Y14i <- rbind(
  subset(Q1Y14, industry_code==10 & own_code==5),
  subset(Q1Y14, industry_code==31212),
  subset(Q1Y14, industry_code==722 & own_code==5),
  subset(Q1Y14, industry_code==72233 & own_code==5))
y14i <- rbind(Q4Y14i,Q3Y14i,Q2Y14i,Q1Y14i)

Q4Y13 <- blsQCEW('Area', year='2013', quarter='4', area='01073')
Q3Y13 <- blsQCEW('Area', year='2013', quarter='3', area='01073')
Q2Y13 <- blsQCEW('Area', year='2013', quarter='2', area='01073')
Q1Y13 <- blsQCEW('Area', year='2013', quarter='1', area='01073')

Q4Y13i <- rbind(
  subset(Q4Y13, industry_code==10 & own_code==5),
  subset(Q4Y13, industry_code==31212),
  subset(Q4Y13, industry_code==722 & own_code==5),
  subset(Q4Y13, industry_code==72233 & own_code==5))
Q3Y13i <- rbind(
  subset(Q3Y13, industry_code==10 & own_code==5),
  subset(Q3Y13, industry_code==31212),
  subset(Q3Y13, industry_code==722 & own_code==5),
  subset(Q3Y13, industry_code==72233 & own_code==5))
Q2Y13i <- rbind(
  subset(Q2Y13, industry_code==10 & own_code==5),
  subset(Q2Y13, industry_code==31212),
  subset(Q2Y13, industry_code==722 & own_code==5),
  subset(Q2Y13, industry_code==72233 & own_code==5))
Q1Y13i <- rbind(
  subset(Q1Y13, industry_code==10 & own_code==5),
  subset(Q1Y13, industry_code==31212),
  subset(Q1Y13, industry_code==722 & own_code==5),
  subset(Q1Y13, industry_code==72233 & own_code==5))
y13i <- rbind(Q4Y13i,Q3Y13i,Q2Y13i,Q1Y13i)

#Data only available until 2012
#2012 was available when I began this project, but the BLS has since taken it down from their website.


#Jeff_county_qcew <- rbind(y16i, y15i, y14i, y13i, y12i)
Jeff_county_qcew <- read.csv("/Users/michaelfinnegan/Documents/Github/BirminghamGentrification/Jeff_cou
drops <- c("agglvl_code", "size_code", "disclosure_code", "lq_disclosure_code", "oty_disclosure_code")
Jeff_county_qcew <- Jeff_county_qcew[ , !(names(Jeff_county_qcew) %in% drops)]
Jeff_county_qcew$yearqtr <- paste(Jeff_county_qcew$year,Jeff_county_qcew$qtr, sep="-")
Jeff_county_qcew$yearqtr <- substr(Jeff_county_qcew$yearqtr,3,6)
```

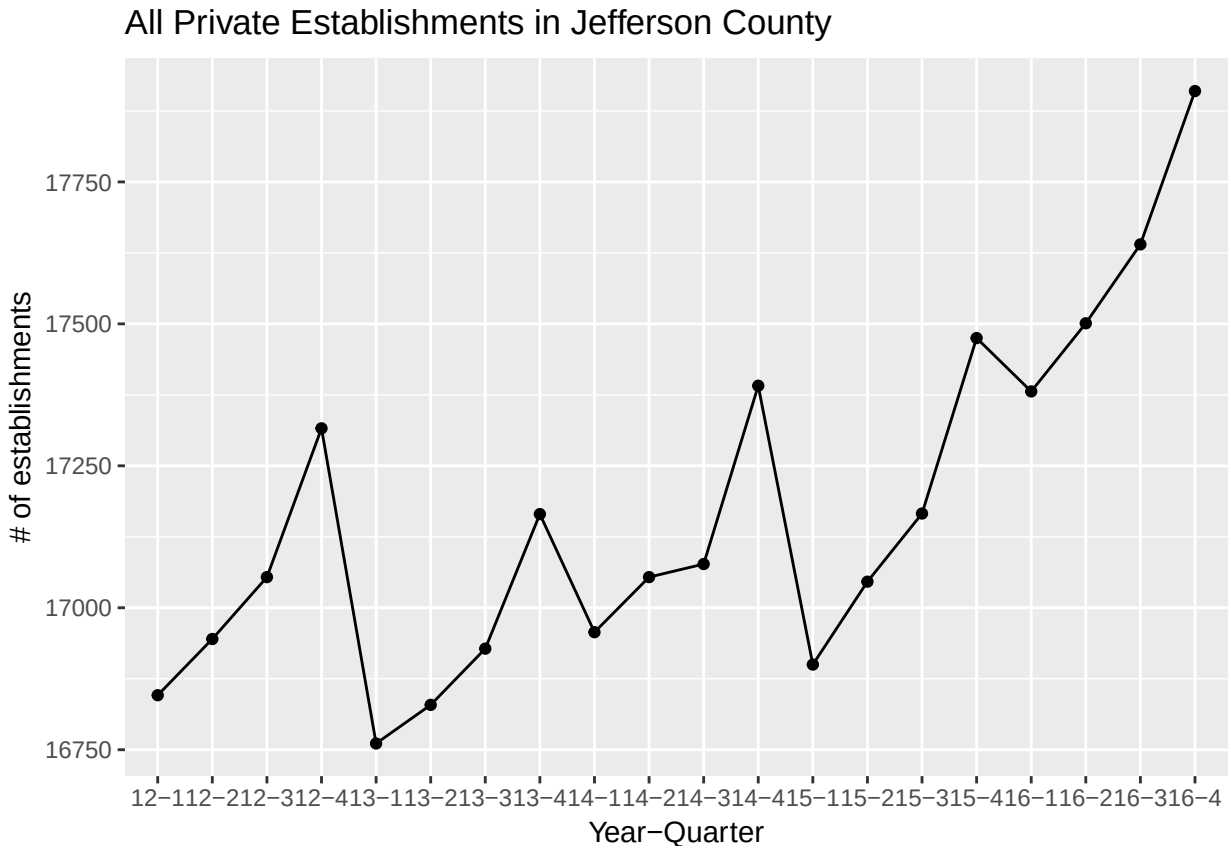
The above block of data simply pulls in QCEW data on Jefferson county for four NAICS industry codes. The first code, i.e. 10, is an aggregate of all NAICS industries in Jefferson County. The following codes, 31212, 722, and 72233, are for breweries, restaurants and bars, and food trucks respectively. It should be noted that the data on food trucks is intrinsically included in the data on all restaurants and bars. While one could subtract the food truck data from restaurants and bars for a more strict comparison, due to the small relative size of food trucks as a percentage of all restaurants and bars, the data has been left unaltered at this point in time.

We are not left with the task of subsetting the data above into seperate industry sets. ##### Create industry subsets

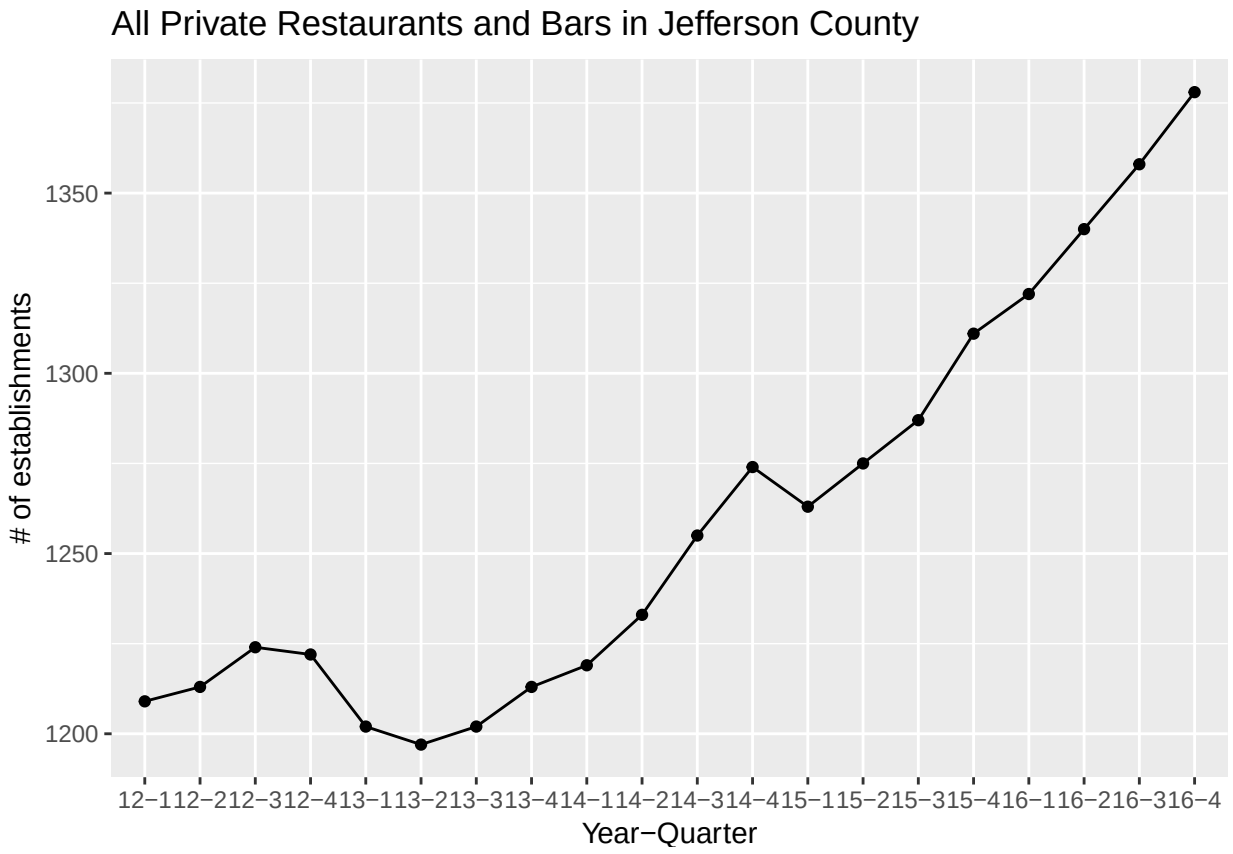
```
#Begin data analysis on Jeff_county_qcew
all <- subset(Jeff_county_qcew, industry_code=="10")
fooddrink <- subset(Jeff_county_qcew, industry_code=="722")
brewery <- subset(Jeff_county_qcew, industry_code=="31212")
foodtruck <- subset(Jeff_county_qcew, industry_code=="72233")
```

Time Series of Establishments

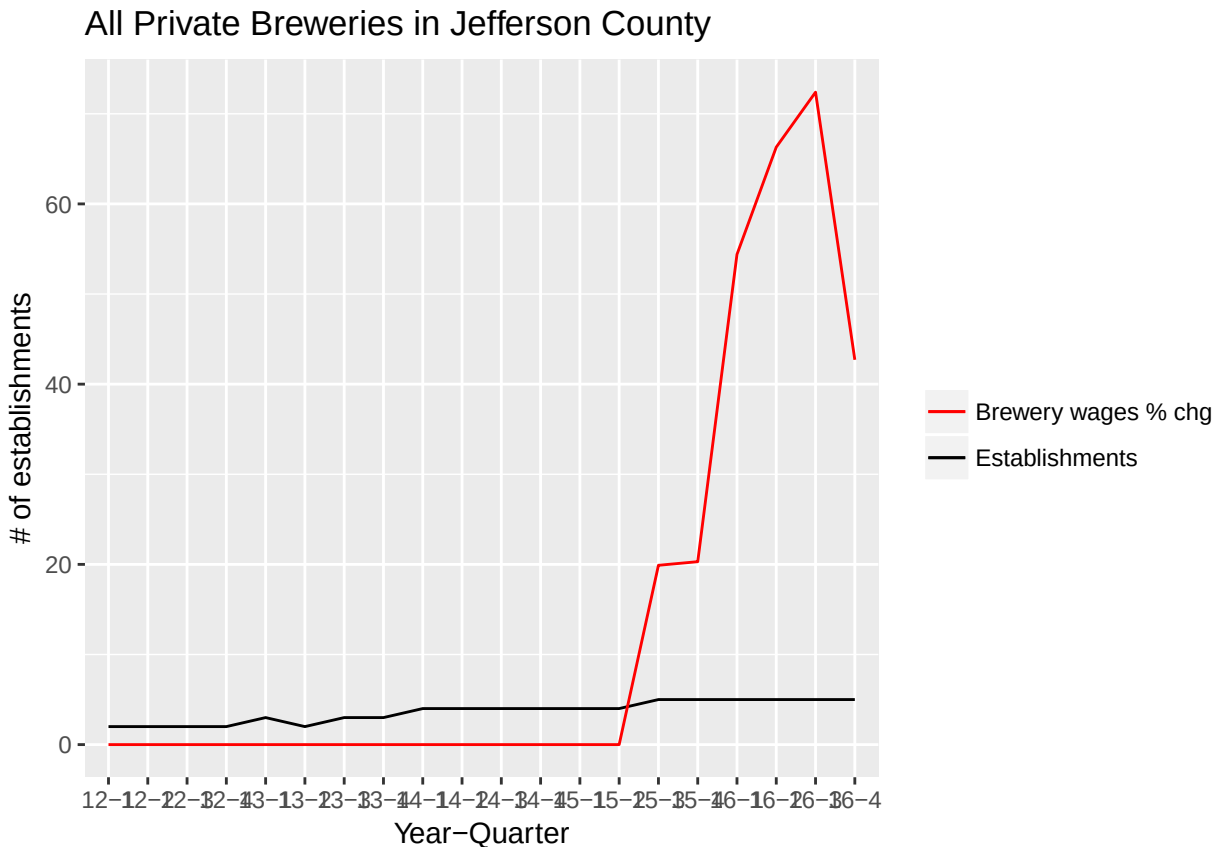
```
ggplot(all, aes(x=yearqtr,qtrly_estabs, group=1)) +
  geom_line() +
  geom_point() +
  xlab("Year-Quarter") +
  ylab("# of establishments") +
ggtitle("All Private Establishments in Jefferson County")
```



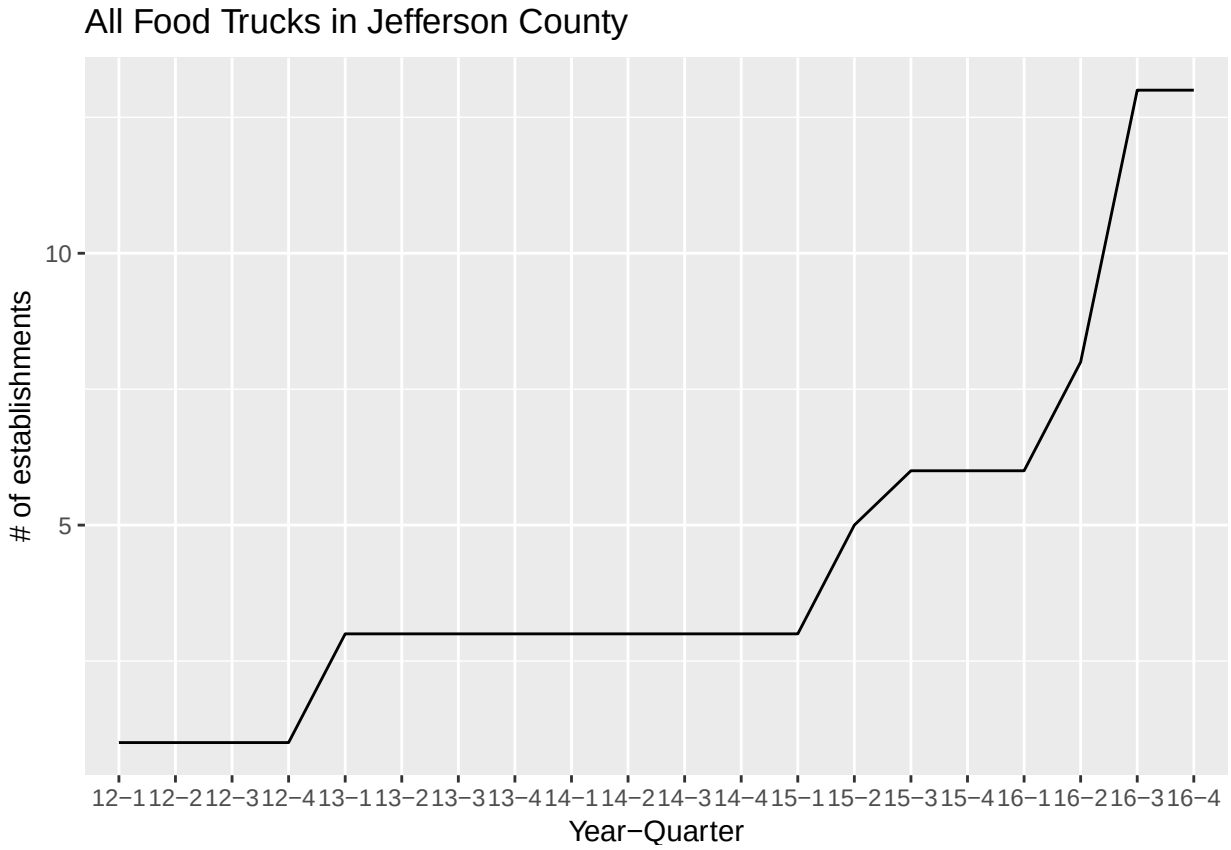

```
ggplot(fooddrink, aes(x=yearqtr,qtrly_estabs, group=1)) +
  geom_line() +
  geom_point() +
  xlab("Year-Quarter") +
  ylab("# of establishments") +
  ggtitle("All Private Restaurants and Bars in Jefferson County")
```




```
ggplot(brewery, aes(x=yearqtr, group=1)) +
  geom_line(aes(y=qtrly_estabs, colour="Establishments")) +
  geom_line(aes(y=oty_total_qtrly_wages_pct_chg, colour="Brewery wages % chg")) +
  scale_colour_manual("",
                      values = c("Establishments"="black", "Brewery wages % chg"="red")) +
  xlab("Year-Quarter") +
  ylab("# of establishments") +
  ggtitle("All Private Breweries in Jefferson County")
```




```
ggplot(foodtruck, aes(x=yearqtr, group=1)) +
  geom_line(aes(y=qtrly_estabs)) +
  xlab("Year-Quarter") +
  ylab("# of establishments") +
  ggtitle("All Food Trucks in Jefferson County")
```



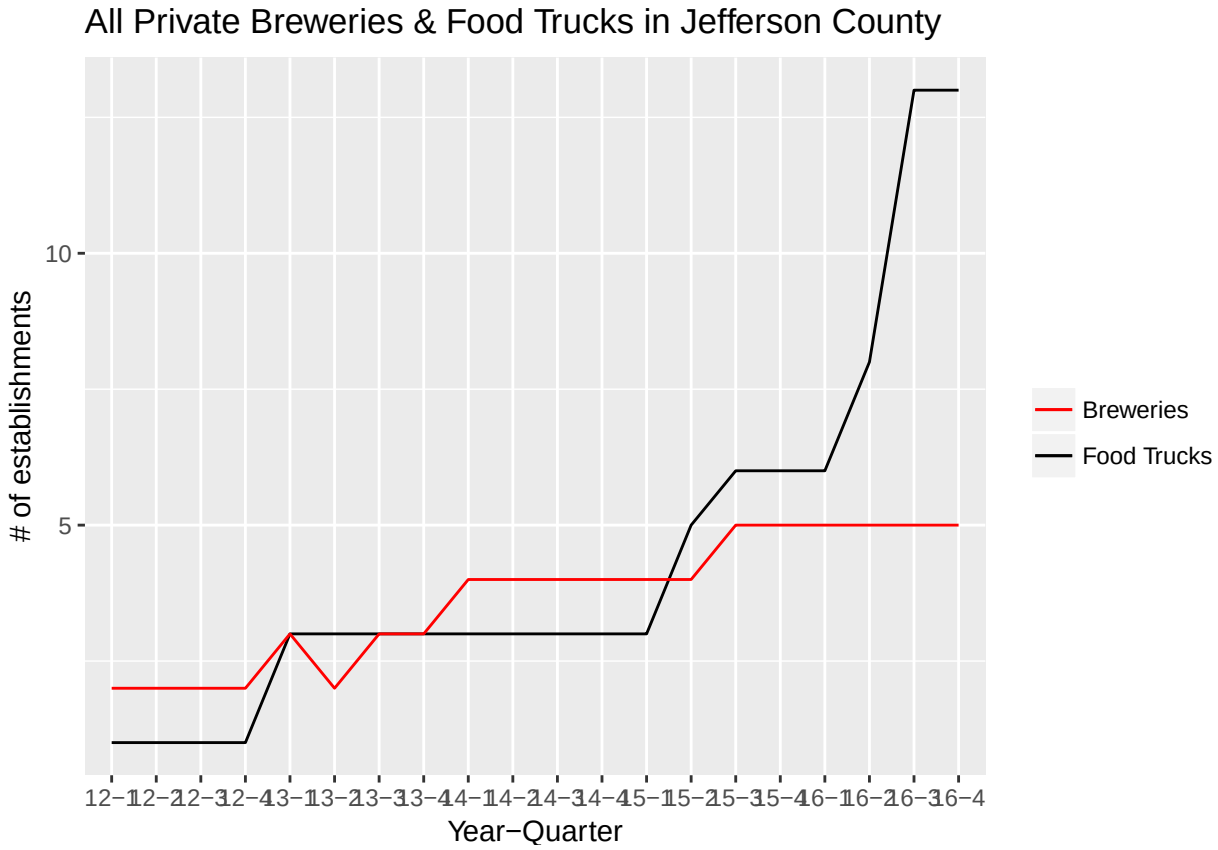
Each of the graphs above show a increase in the number of establishments. The second to last graph of breweries included the over the year percentage change in brewery wages, which unfortunately data only began being recorded for in 2014. The inclusion of this variable was to see if wages have risen as the number of breweries have also risen, which appears to be the case.

Relationship between Breweries and Food Trucks

We will now examine a possible relationship between breweries and food trucks. Intuitively, we assume that many of the same people that are visiting food trucks are also visiting breweries. In fact, many of the breweries in Birmingham allow food trucks to park on the premise of the brewery, such that customers are able to legally get food without having to abandon the beers they purchased while in the brewery.

```
foodtruck$brewery_qtrly_estabs <-brewery$qtrly_estabs

ggplot(foodtruck, aes(x=yearqtr, group=1)) +
  geom_line(aes(y=qtrly_estabs, colour="Food Trucks")) +
  geom_line(aes(y=brewery_qtrly_estabs, colour="Breweries")) +
  scale_colour_manual("",
                      values = c("Food Trucks"="black", "Breweries"="red")) +
  xlab("Year-Quarter") +
  ylab("# of establishments") +
  ggtitle("All Private Breweries & Food Trucks in Jefferson County")
```



The graph above leads us to believe the relationship between breweries and food trucks exist. That we know for a fact that the food trucks will literally park on the property of the brewery is evidence enough of a relationship between the two. However, the statistical significance as we will see below may alert us to interesting effects.

For example, in any modeling effort we will certainly want to include and interaction between food trucks and breweries. Furthermore, the presence of the two in combination could lead to an important negative externality on certain neighborhoods: the creation of a food desert (an issue with which Birmingham has long struggled with).

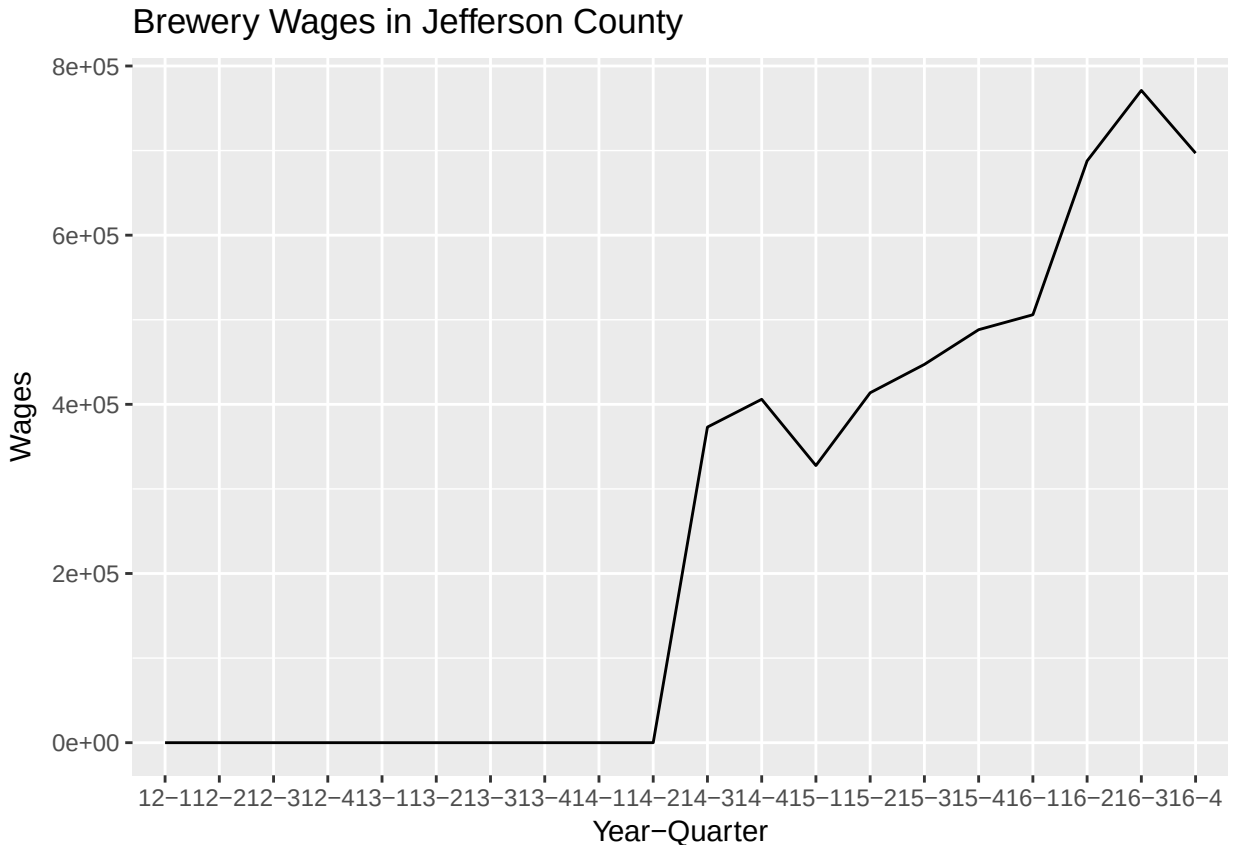
Correlation test

```
cor.test(foodtruck$qtrly_estabs, brewery$qtrly_estabs)
```

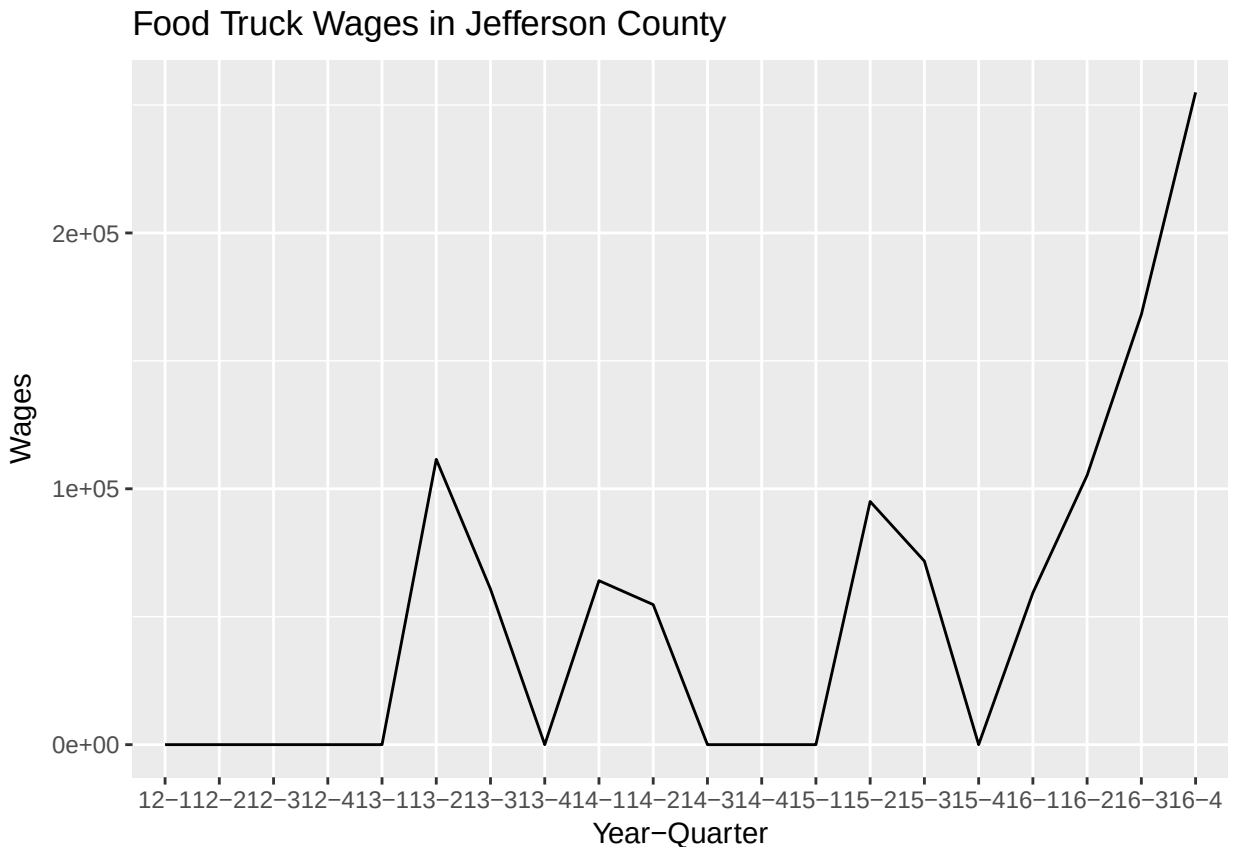
```
##
##  Pearson's product-moment correlation
##
## data:  foodtruck$qtrly_estabs and brewery$qtrly_estabs
## t = 4.9711, df = 18, p-value = 9.884e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4795223 0.9001645
## sample estimates:
##      cor
## 0.760637
```

A look at wages

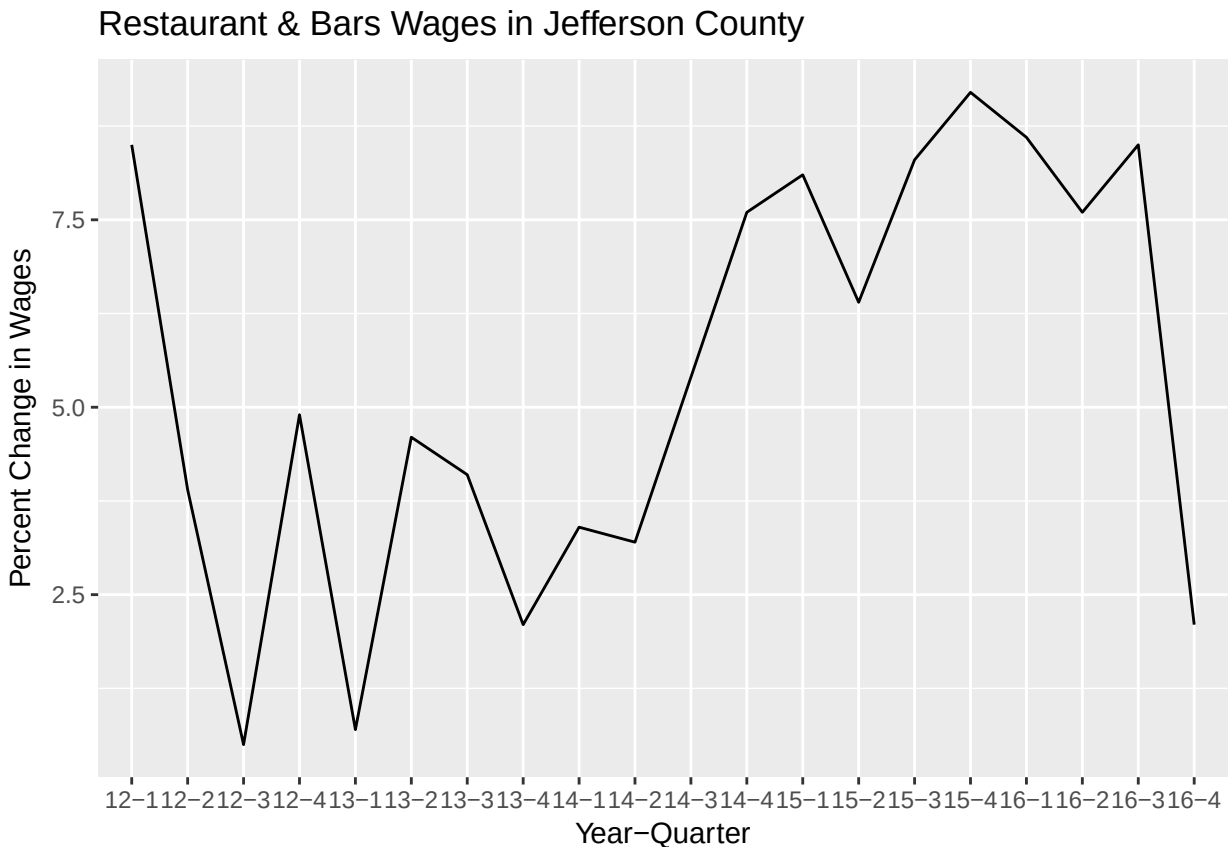
```
ggplot(brewery, aes(x=yearqtr, group=1)) +
  geom_line(aes(y=total_qtrly_wages)) +
  xlab("Year-Quarter") +
  ylab("Wages") +
  ggtitle("Brewery Wages in Jefferson County")
```



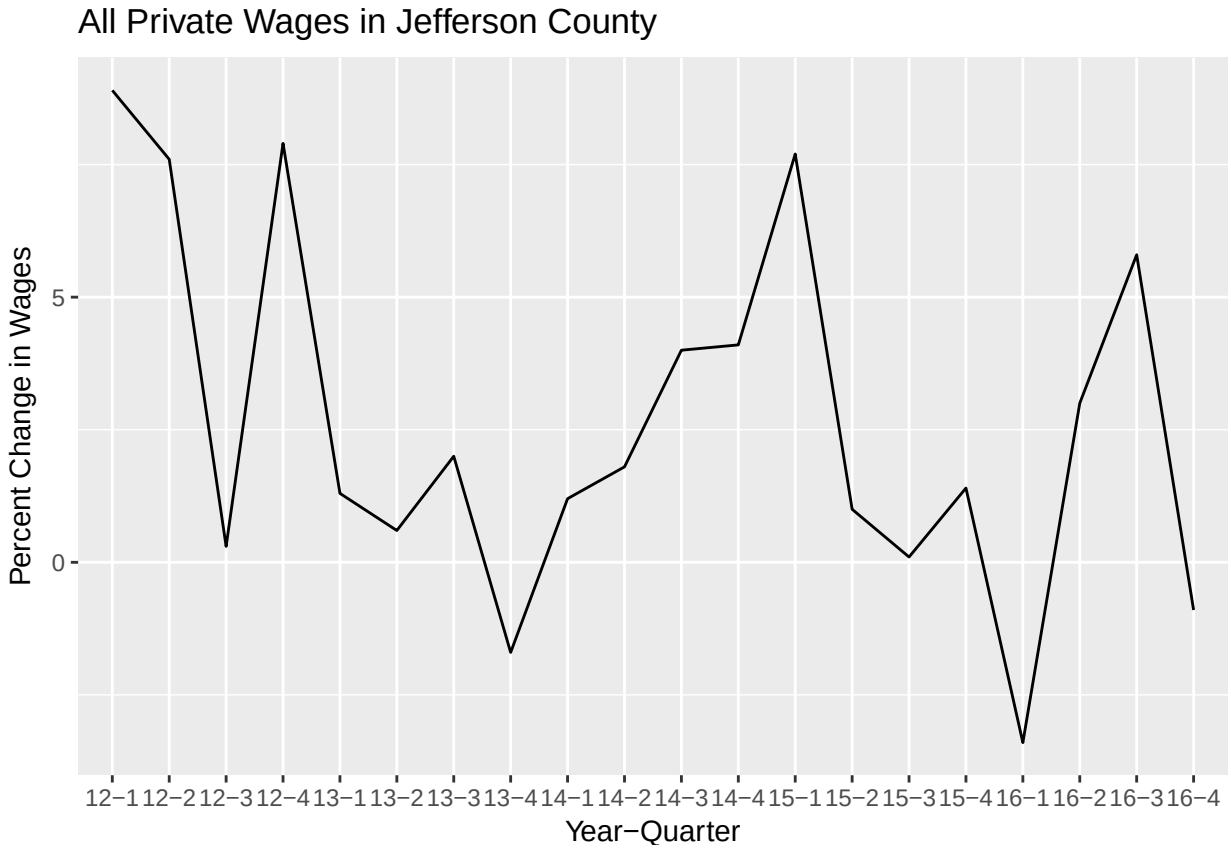

```
ggplot(foodtruck, aes(x=yearqtr, group=1)) +
  geom_line(aes(y=total_qtrly_wages)) +
  xlab("Year-Quarter") +
  ylab("Wages") +
  ggtitle("Food Truck Wages in Jefferson County")
```




```
ggplot(fooddrink, aes(x=yearqtr, group=1)) +
  geom_line(aes(y=oty_total_qtrly_wages_pct_chg)) +
  xlab("Year-Quarter") +
  ylab("Percent Change in Wages") +
  ggtitle("Restaurant & Bars Wages in Jefferson County")
```




```
ggplot(all, aes(x=yearqtr, group=1)) +
  geom_line(aes(y=oty_total_qtrly_wages_pct_chg)) +
  xlab("Year-Quarter") +
  ylab("Percent Change in Wages") +
  ggtitle("All Private Wages in Jefferson County")
```

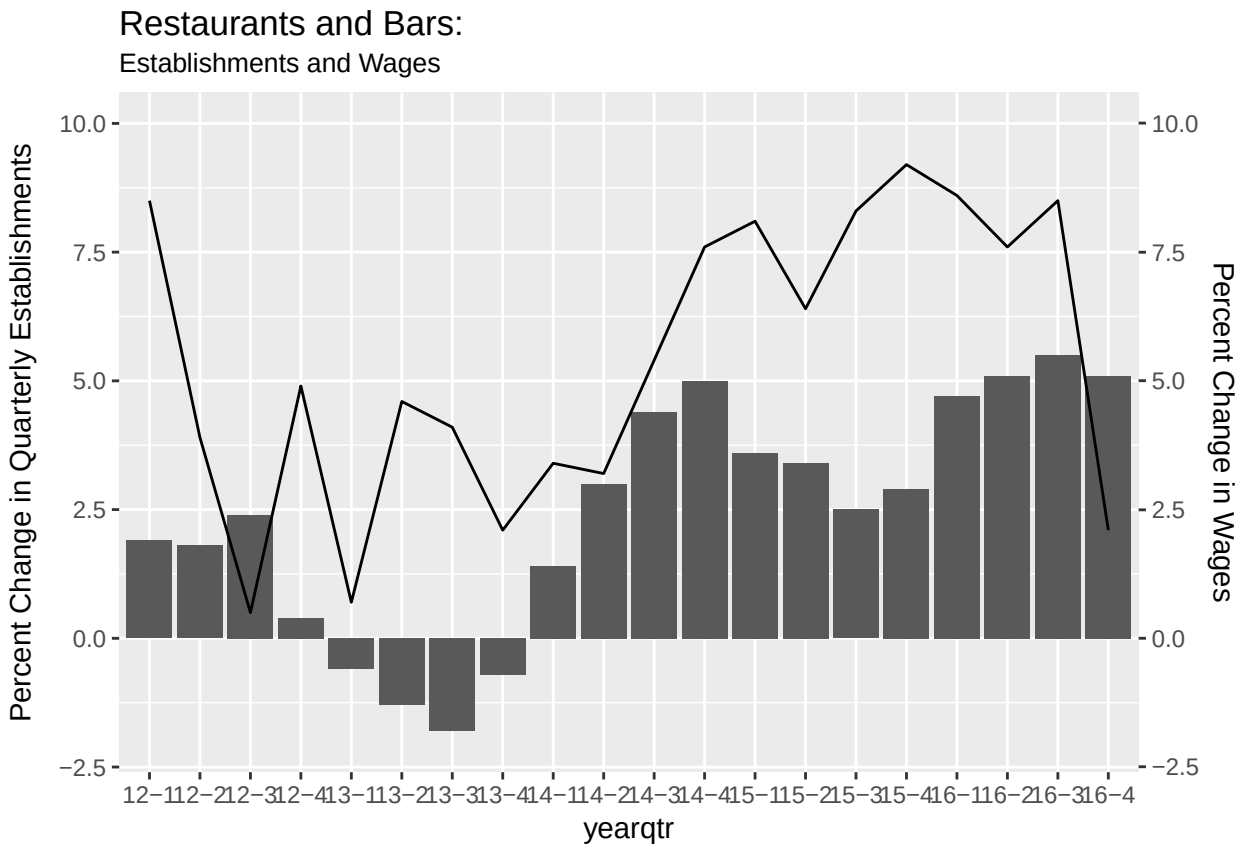


Download necessary packages for extended analysis on wages

```
library(pipeR)
library(readr)
library(lubridate)
```

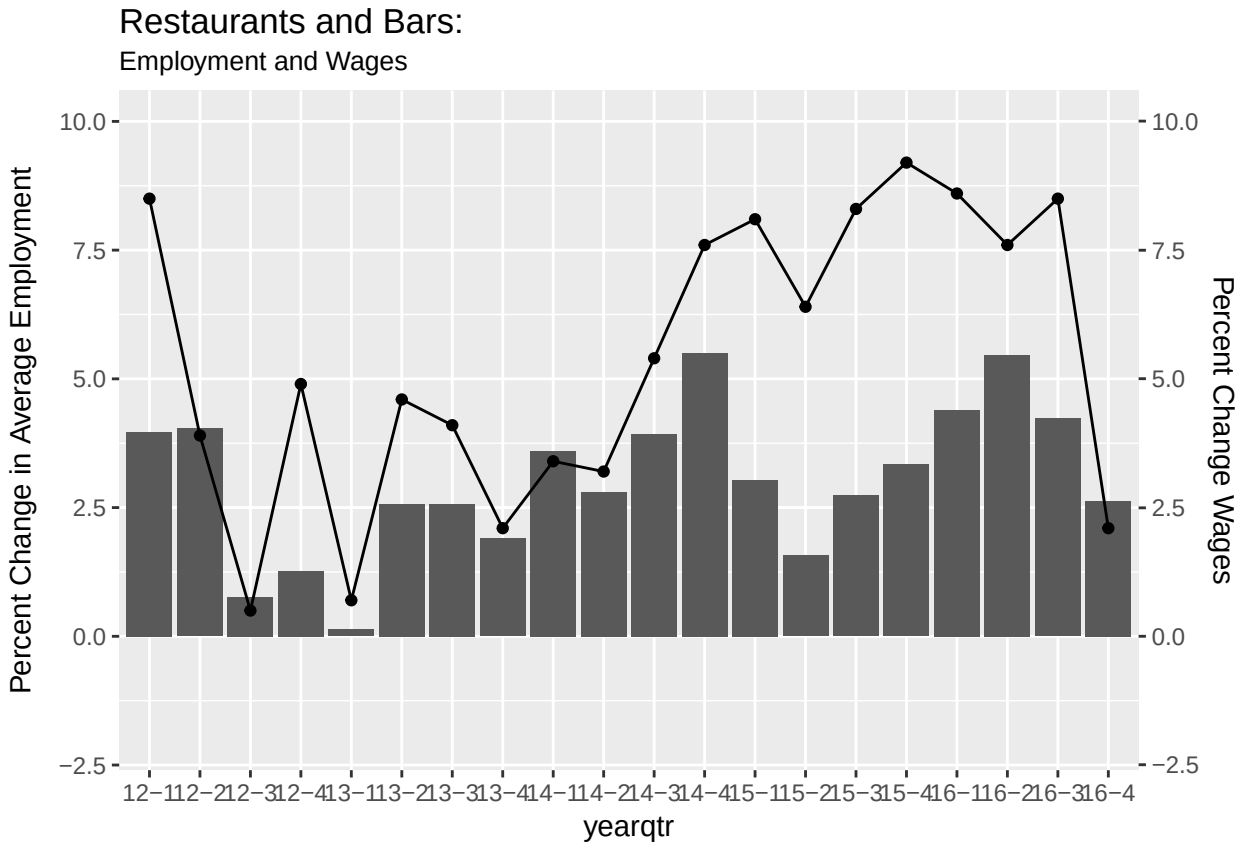
Looking at wages and employment together

```
plot1 <- fooddrink %>>% ggplot() +
  geom_bar(mapping = aes(x = yearqtr, y = oty_qtrly_estabs_pct_chg),
           stat = "identity") +
  #geom_point(mapping = aes(x = yearqtr, y = total_qtrly_wages)) +
  geom_line(mapping = aes(x = yearqtr, y = oty_total_qtrly_wages_pct_chg, group=1)) +
  scale_y_continuous(name = expression("Percent Change in Quarterly Establishments"),
                     limits = c(-2, 10))
plot2 <- plot1 %+% scale_y_continuous(name =
                    expression("Percent Change in Quarterly Establishments"),
                    sec.axis = sec_axis(~ .,
                        name = "Percent Change in Wages"), limits = c(-2, 10))+
                  labs(title="Restaurants and Bars:",
                       subtitle="Establishments and Wages")
plot2
```




```
fooddrink$oty_avg_emplvl_pct_chg <- rowMeans(subset(fooddrink, select =
                                    c(oty_month1_emplvl_pct_chg,
                                      oty_month2_emplvl_pct_chg,
                                      oty_month3_emplvl_pct_chg)))

plo3 <- fooddrink %>>% ggplot() +
  geom_bar(mapping = aes(x = yearqtr, y = oty_avg_emplvl_pct_chg), stat="identity") +
  geom_point(mapping = aes(x = yearqtr, y = oty_total_qtrly_wages_pct_chg)) +
  geom_line(mapping = aes(x = yearqtr, y = oty_total_qtrly_wages_pct_chg, group=1)) +
  scale_y_continuous(name = expression("Percent Change in Average Employment"),
                     limits = c(-2, 10))
plot4 <- plo3 %+% scale_y_continuous(name =
                                       expression("Percent Change in Average Employment"),
                                     sec.axis = sec_axis(~ .,
                                                         name = "Percent Change Wages"), limits = c(-2,
  labs(title="Restaurants and Bars:",
       subtitle="Employment and Wages")
plot4
```



The two graphs above give us the idea that wages are rising with the growth within Birmingham. It is important to note that the percent changes in wage are calculated from nominal, not real wages.

Another important note from all of the analysis on of the QCEW data is that there is a lot of seasonality as seen in the large increase and decrease on either side of the fourth quarter of every year. We could proceed with accounting for seasonality with the stl function in the loess package.