Sean McLean

ALY 6110

Module 4 Assignment

Practice With Spark

Introduction

This assignment focuses on solving business questions that pertain to two datasets imported into the Sparkly R library. One dataset looks at annual house price indexes with five-digit zip codes, and the question being addressed is how the HPI differs between the 1990 base and 2000 base. The other dataset looks at median home values from the company Zillow, and the question being asked pertains to what parts of the country have the highest peak values historically. To better understand the relationships between the variables of the datasets, thorough analysis and visualizations are conducted that will provide the necessary insights.

Analysis & Results

HPI Five Digit Dataset

After installing the necessary packages and libraries applied from the module assignment reference, the dataset was imported and copied to Spark. To verify that the dataset was properly downloaded, the head function was run to show the top five observations for each variable. The usage of summary statistics for the dataset provided some insight into the size of the dataset and what types of variables were included. Through functions like str() and summary(), a total of six variables and 583,050 observations are contained in the dataset.

Using the DBI library provided from the module reference and the dplyr library, a sample is used from the dataset and a plot is created to show the relationship between the variables that show the HPI of 1990 and 2000 (See Appendix A). The results show a positive linear relationship as both variables in 1990 and 2000 show the values increasing over time. An analysis where descriptive statistics are employed by creating a new variable that calculates the average housing price index for the 1990 base and 2000 base. A bar plot is made using the ggplot() function that where the heights of the bars represent the values of the specified variables (See Appendix B). The plot shows that among the two variables the 1990 base is higher than the 2000 base in average HPI.

A correlation analysis is done by first converting all the variables to numeric and then calculating the matrix to show the strength of the relationships between the variables. The plot of the matrix shows stronger relationships than weak ones, but there are also many variables that do not show any connection (See Appendix C). A dbplot histogram showing the number of observations per bin that encompasses a three-year period is executed. The results of the plot show that the number of houses in the dataset increases over the 40-year period, with the most houses by a considerable amount in the most recent bin that includes the years 2018-2020 (See Appendix D).

Zillow Dataset

After installing the necessary packages and libraries applied from the module assignment reference, the dataset was imported and copied to Spark. To verify that the dataset was properly downloaded, the head function was operated to display the top five observations for each variable. Adding summary statistics for the dataset provided clarity into the dataset's size and what types of variables were included. Through functions like str() and summary(), a total of 15 variables and 30,134 observations are contained in the dataset.

Using the DBI library provided from the module reference and the dplyr library, linear regression analysis is executed between the variables that show the percentage fall from the peak and the peak of the Zillow house index. The results of the coefficients in the variables are –2.48

for the intercept and –3.41 for the peak Zillow house index. A scatterplot is formulated to visualize predictions versus actual values and provides a range that shows the upper echelon of home prices. The plot indicates that the smaller the house price the closer the percentage falls from peak is closer to zero (See Appendix E).

A descriptive analysis that includes aggregation and ranking addresses the question for the dataset and uses the variables pertaining to each state and the peak Zillow house value index. The results are displayed by a bar chart using the ggplot() function which gives the top 10 states with the highest peak ZHVI. The top state is California by a considerable margin followed by New York and Massachusetts (See Appendix F). A time series analysis studying the relationships between the states and the average ZHVI is conducted by first creating a new column called 'PeakYear' from calculating each year from the 'PeakMonth' variable. The ZHVI mean of each state and year is then computed, and the results are visualized with a line plot showing how states fare with their average ZHVI per year. Most states remain low over time with some recent jumps by a handful of states in the last decade (See Appendix G).

Insights

HPI Five Digit Dataset

Looking at the business question that pertains to the difference between the house price indexes between the 1990 base and the 2000 base, some of the plots show that the 2000 base is lower than the 1990 base. When applying the bar plot of total houses in the dataset per year, there are more new houses than older ones. This could indicate that the HPI is lower in 2000 because there were more available homes, and the housing market could have saturated to the point that prices were lower. The first plot showing a sample the relationship between both variables does show a positive linear relationship with prices increasing, but there are only a few data points that stick out from most of the sample. There could be a small margin of error overall between the two variables with some plots showing one direction and results showing the other direction.

The matrix shows the strongest, weakest, and neutral relationships among variables, and the strongest were based off the year of the house and the HPI’s overall and from both the 1990 base and 2000 base. The values of homes therefore weigh on how old the home is and how it appreciates over time. The annual change variable also had a strong relationship with zip codes of the homes and the HPI, so this could have been the result of a neighborhood being gentrified around a certain period which had an impact on home values in the area. There were some variables with almost zero relationship between them, and the negative relationships were just slightly below zero. The conclusion of the matrix is that there are no relationships that were heavily impacted in a negative way. From these insights I can conclude that there are only slight differences between the two HPI bases in 1990 and 2000.

Zillow Dataset

The negative slope coefficient from the linear regression indicates a slight inverse relationship between PeakZHVI and PctFallFromPeak, meaning that as PeakZHVI increases, PctFallFromPeak decreases marginally. The minimal magnitude of the slope suggests that PeakZHVI has an extremely limited effect on PctFallFromPeak, implying that other factors not included in this model might have a more significant impact. These results seem to be in line with what the plot of the linear regression is showing. The results from the statistics analysis and then bar plot show the top 10 states by peak Zillow housing value index are mainly all states from the west coast and east coast, suggesting that people are more likely to seek homes in a state by the coast. Besides Ohio, the states are all either in the Northeast, Southeast, and western part regions of the United States.

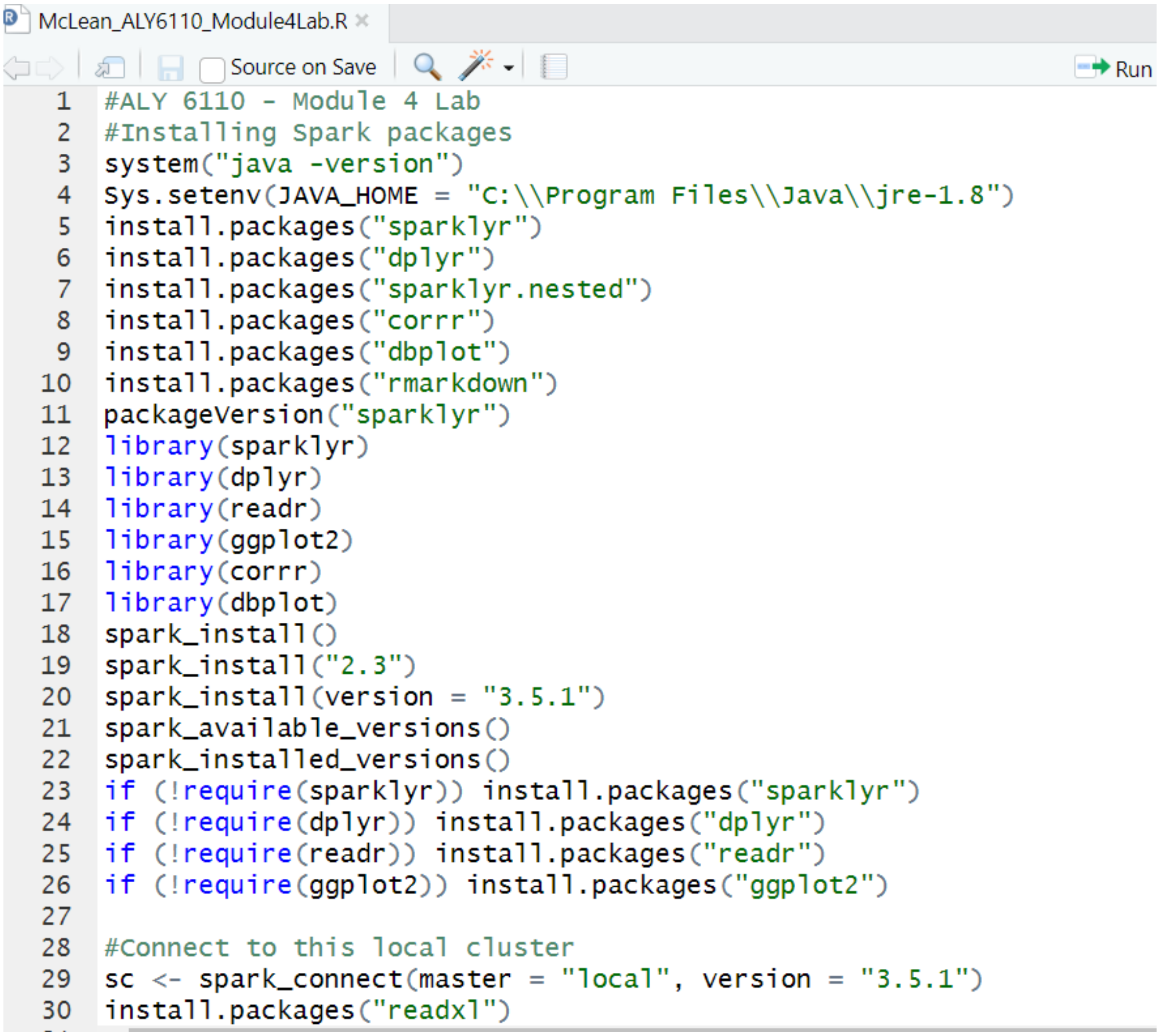
The time series analysis results with the calculations showing in the line plot entail many states with just moderate increases from the starting point of 1995 to 2020. The few states that saw larger increases began just around 2008 with either Florida or Georgia being the one outlier with a very substantial increase in average ZHVI for the year 2017. While it is difficult to see which states exactly had increases, the time it happened might pertain to the economic collapse that was connected to real estate and prices of homes. The one year where the average value of homes drastically increased in Florida and Georgia could indicate that there was a remarkably high demand to move there at the time which resulted in the value of homes increasing overall. From these insights I can conclude that the top states by value over time are based off geography and from certain events happening that affected the house values.

References

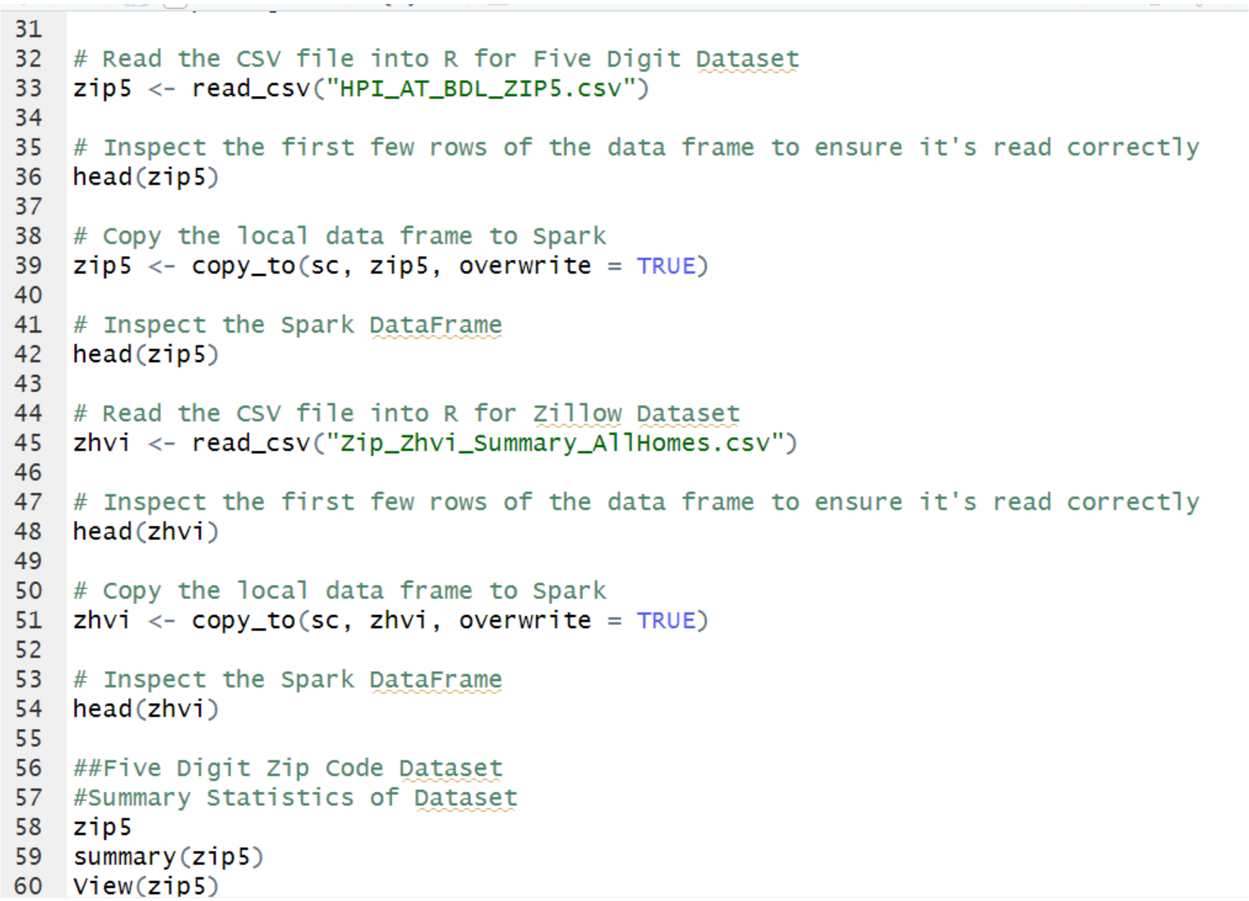
Luraschi, J., Kuo, K. & Ruiz, E. (2019, Nov. 19th). *Mastering Spark with R: The Complete Guide to Large-Scale Analysis and Modeling 1st Edition.* O'Reilly Media.

OpenAI. (2024). *ChatGPT* (June 12 version). https://chatgpt.com/

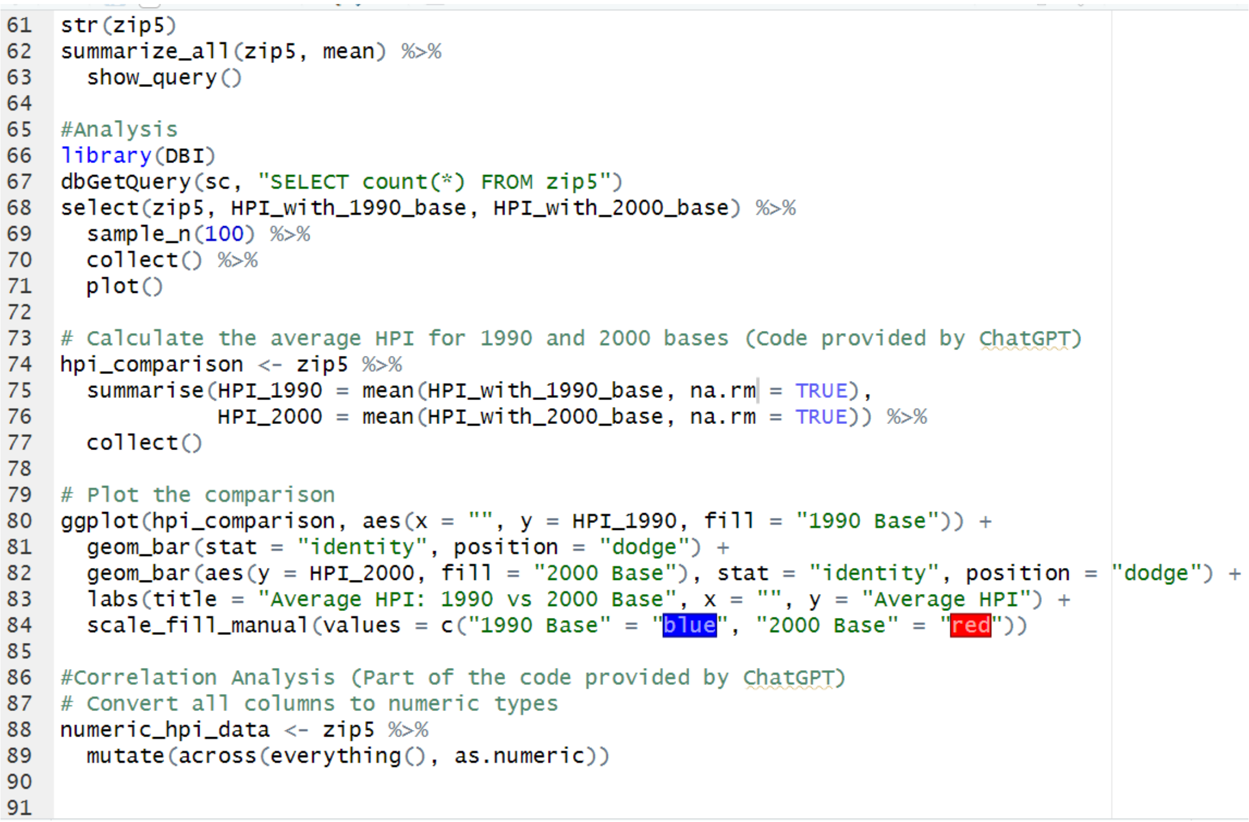
Appendix



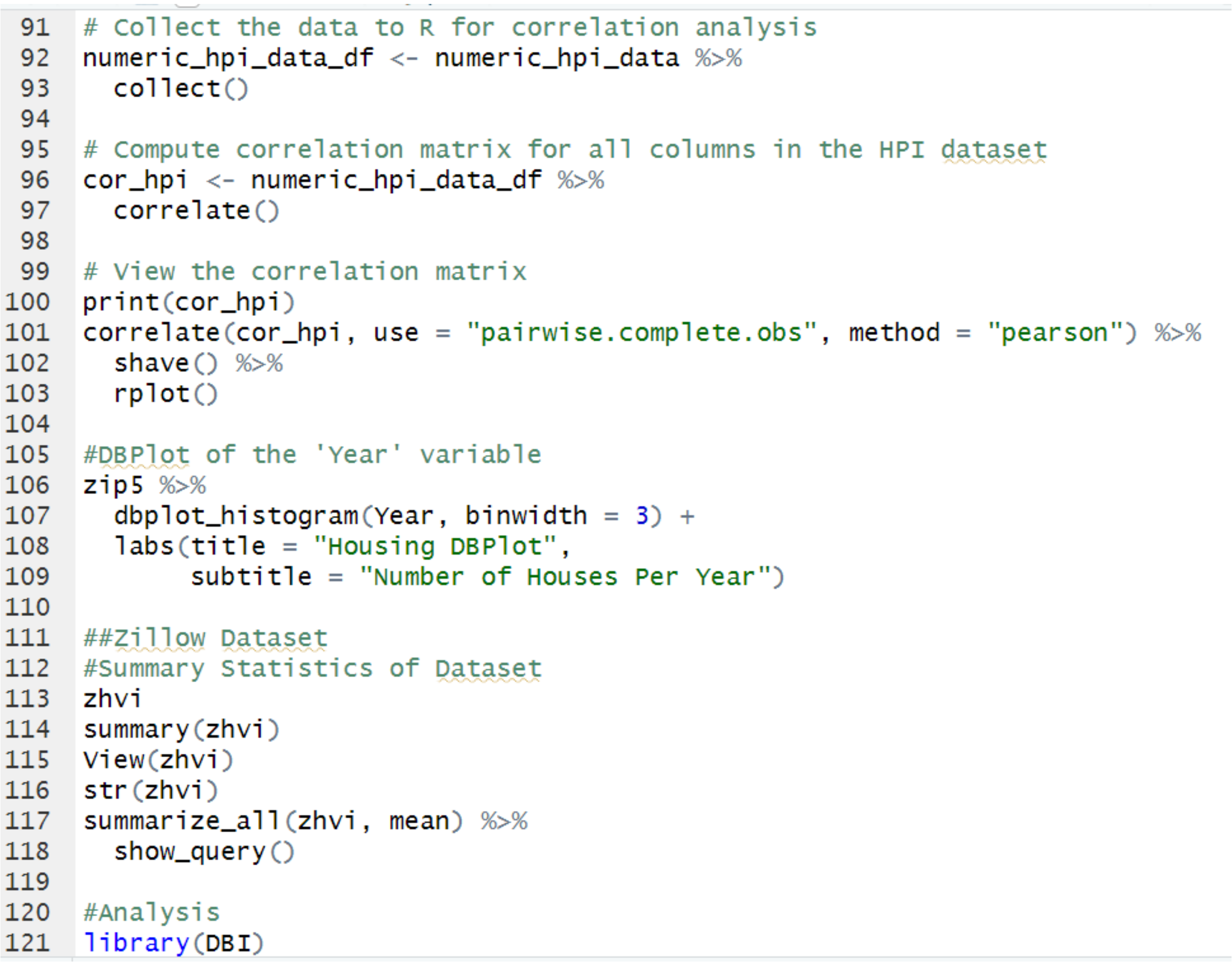
```r
#ALY 6110 - Module 4 Lab
#Installing Spark packages
system("java -version")
Sys.setenv(JAVA_HOME = "C:\\Program Files\\Java\\jre-1.8")
install.packages("sparklyr")
install.packages("dplyr")
install.packages("sparklyr.nested")
install.packages("corrr")
install.packages("dbplot")
install.packages("rmarkdown")
packageVersion("sparklyr")
library(sparklyr)
library(dplyr)
library(readr)
library(ggplot2)
library(corrr)
library(dbplot)
spark_install()
spark_install("2.3")
spark_install(version = "3.5.1")
spark_available_versions()
spark_installed_versions()
if (!require(sparklyr)) install.packages("sparklyr")
if (!require(dplyr)) install.packages("dplyr")
if (!require(readr)) install.packages("readr")
if (!require(ggplot2)) install.packages("ggplot2")

#Connect to this local cluster
sc <- spark_connect(master = "local", version = "3.5.1")
install.packages("readxl")
```



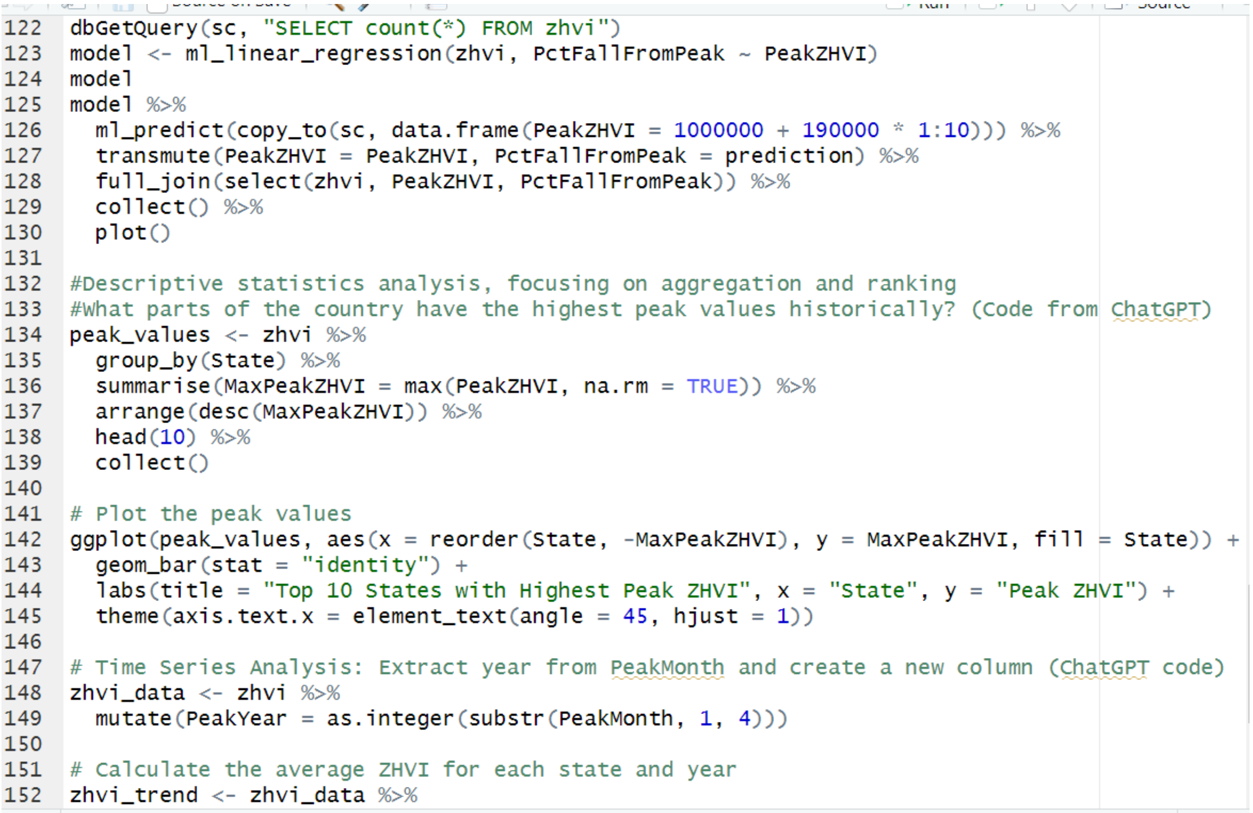
```r

# Read the CSV file into R for Five Digit Dataset
zip5 <- read_csv("HPI_AT_BDL_ZIP5.csv")

# Inspect the first few rows of the data frame to ensure it's read correctly
head(zip5)

# Copy the local data frame to Spark
zip5 <- copy_to(sc, zip5, overwrite = TRUE)

# Inspect the Spark DataFrame
head(zip5)

# Read the CSV file into R for Zillow Dataset
zhvi <- read_csv("Zip_Zhvi_Summary_AllHomes.csv")

# Inspect the first few rows of the data frame to ensure it's read correctly
head(zhvi)

# Copy the local data frame to Spark
zhvi <- copy_to(sc, zhvi, overwrite = TRUE)

# Inspect the Spark DataFrame
head(zhvi)

##Five Digit Zip Code Dataset
#Summary Statistics of Dataset
zip5
summary(zip5)
View(zip5)
```



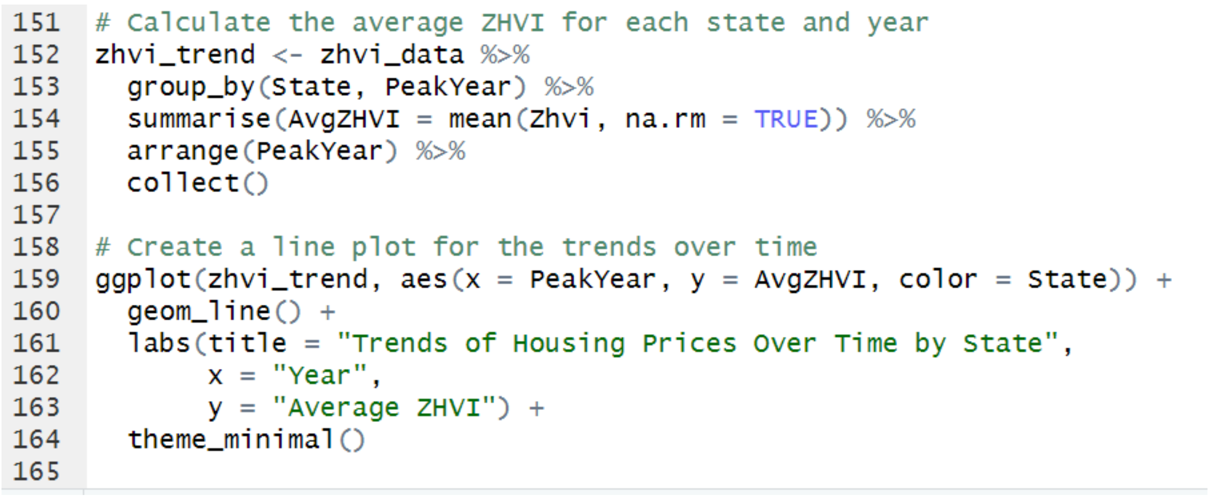
```r
str(zip5)
summarize_all(zip5, mean) %>%
  show_query()

#Analysis
library(DBI)
dbGetQuery(sc, "SELECT count(*) FROM zip5")
select(zip5, HPI_with_1990_base, HPI_with_2000_base) %>%
  sample_n(100) %>%
  collect() %>%
  plot()

# Calculate the average HPI for 1990 and 2000 bases (Code provided by ChatGPT)
hpi_comparison <- zip5 %>%
  summarise(HPI_1990 = mean(HPI_with_1990_base, na.rm = TRUE),
            HPI_2000 = mean(HPI_with_2000_base, na.rm = TRUE)) %>%
  collect()

# Plot the comparison
ggplot(hpi_comparison, aes(x = "", y = HPI_1990, fill = "1990 Base")) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_bar(aes(y = HPI_2000, fill = "2000 Base"), stat = "identity", position = "dodge") +
  labs(title = "Average HPI: 1990 vs 2000 Base", x = "", y = "Average HPI") +
  scale_fill_manual(values = c("1990 Base" = "blue", "2000 Base" = "red"))

#Correlation Analysis (Part of the code provided by ChatGPT)
# Convert all columns to numeric types
numeric_hpi_data <- zip5 %>%
  mutate(across(everything(), as.numeric))


```



```r
# Collect the data to R for correlation analysis
numeric_hpi_data_df <- numeric_hpi_data %>%
  collect()

# Compute correlation matrix for all columns in the HPI dataset
cor_hpi <- numeric_hpi_data_df %>%
  correlate()

# View the correlation matrix
print(cor_hpi)
correlate(cor_hpi, use = "pairwise.complete.obs", method = "pearson") %>%
  shave() %>%
  rplot()

#DBPlot of the 'Year' variable
zip5 %>%
  dbplot_histogram(Year, binwidth = 3) +
  labs(title = "Housing DBPlot",
       subtitle = "Number of Houses Per Year")

##Zillow Dataset
#Summary Statistics of Dataset
zhvi
summary(zhvi)
View(zhvi)
str(zhvi)
summarize_all(zhvi, mean) %>%
  show_query()

#Analysis
library(DBI)
```

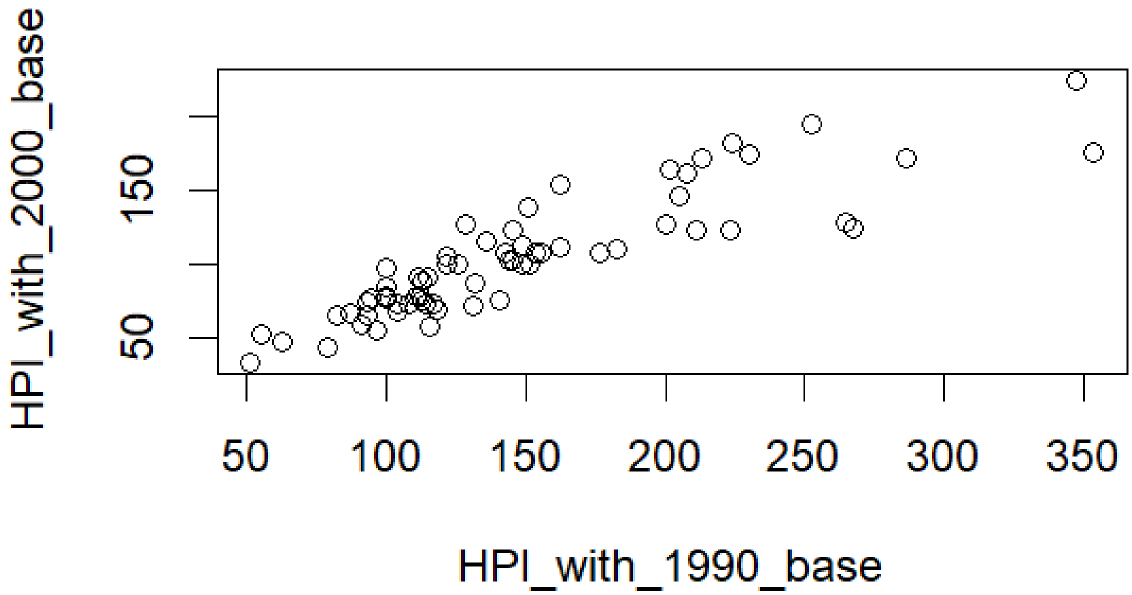


```r
dbGetQuery(sc, "SELECT count(*) FROM zhvi")
model <- ml_linear_regression(zhvi, PctFallFromPeak ~ PeakZHVI)
model
model %>%
  ml_predict(copy_to(sc, data.frame(PeakZHVI = 1000000 + 190000 * 1:10))) %>%
  transmute(PeakZHVI = PeakZHVI, PctFallFromPeak = prediction) %>%
  full_join(select(zhvi, PeakZHVI, PctFallFromPeak)) %>%
  collect() %>%
  plot()

#Descriptive statistics analysis, focusing on aggregation and ranking
#What parts of the country have the highest peak values historically? (Code from ChatGPT)
peak_values <- zhvi %>%
  group_by(State) %>%
  summarise(MaxPeakZHVI = max(PeakZHVI, na.rm = TRUE)) %>%
  arrange(desc(MaxPeakZHVI)) %>%
  head(10) %>%
  collect()

# Plot the peak values
ggplot(peak_values, aes(x = reorder(State, -MaxPeakZHVI), y = MaxPeakZHVI, fill = State)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 10 States with Highest Peak ZHVI", x = "State", y = "Peak ZHVI") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Time Series Analysis: Extract year from PeakMonth and create a new column (ChatGPT code)
zhvi_data <- zhvi %>%
  mutate(PeakYear = as.integer(substr(PeakMonth, 1, 4)))

# Calculate the average ZHVI for each state and year
zhvi_trend <- zhvi_data %>%
```

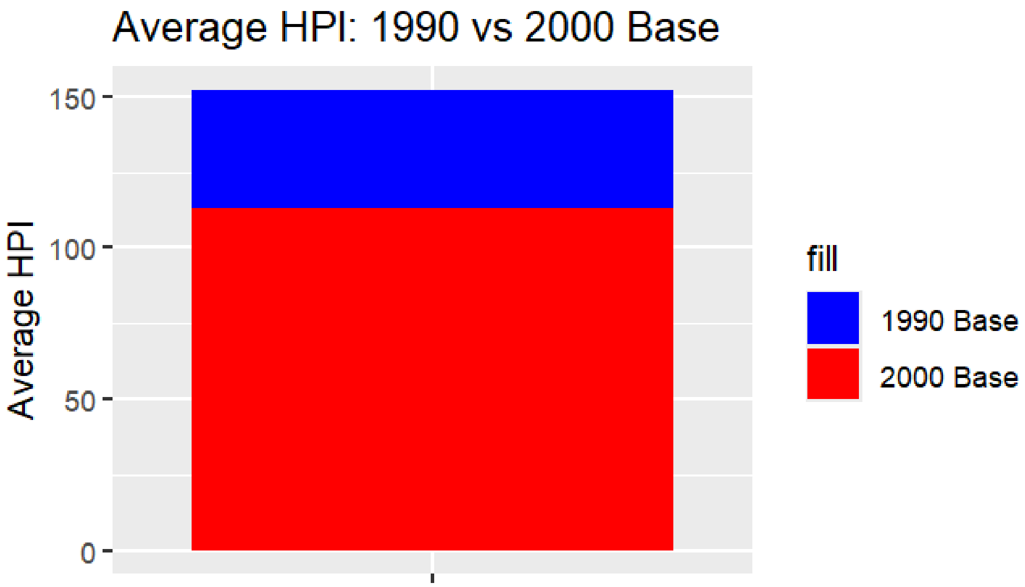


```r
# Calculate the average ZHVI for each state and year
zhvi_trend <- zhvi_data %>%
  group_by(State, PeakYear) %>%
  summarise(AvgZHVI = mean(Zhvi, na.rm = TRUE)) %>%
  arrange(PeakYear) %>%
  collect()

# Create a line plot for the trends over time
ggplot(zhvi_trend, aes(x = PeakYear, y = AvgZHVI, color = State)) +
  geom_line() +
  labs(title = "Trends of Housing Prices Over Time by State",
       x = "Year",
       y = "Average ZHVI") +
  theme_minimal()
```

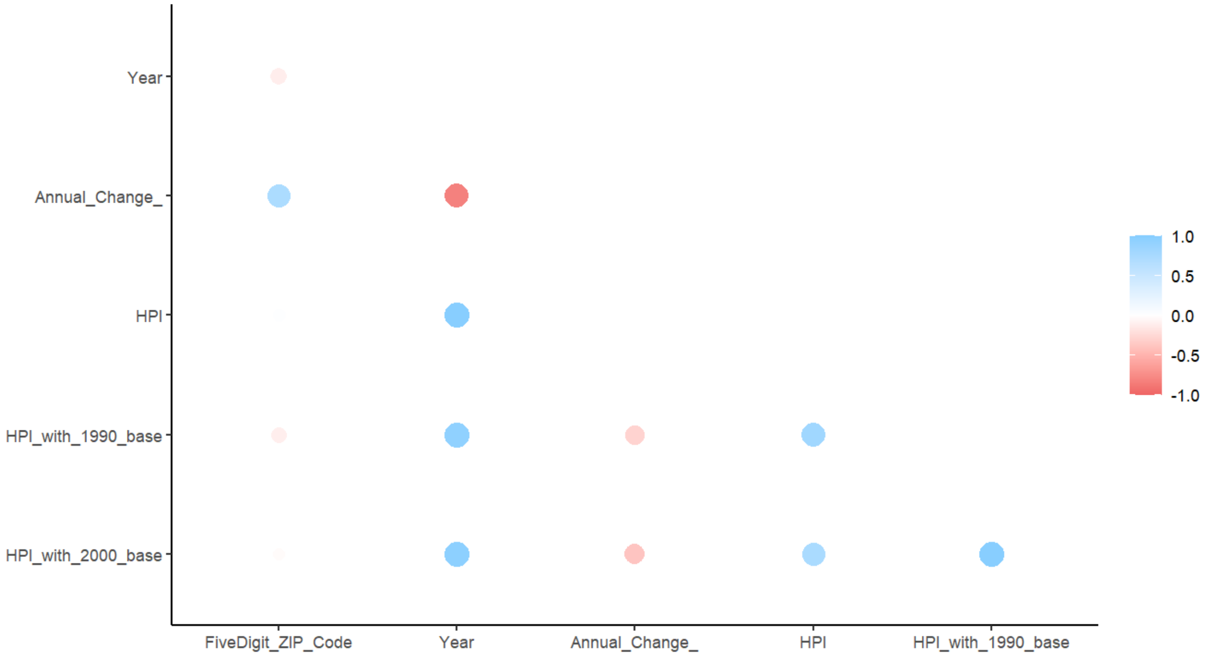
Appendix A



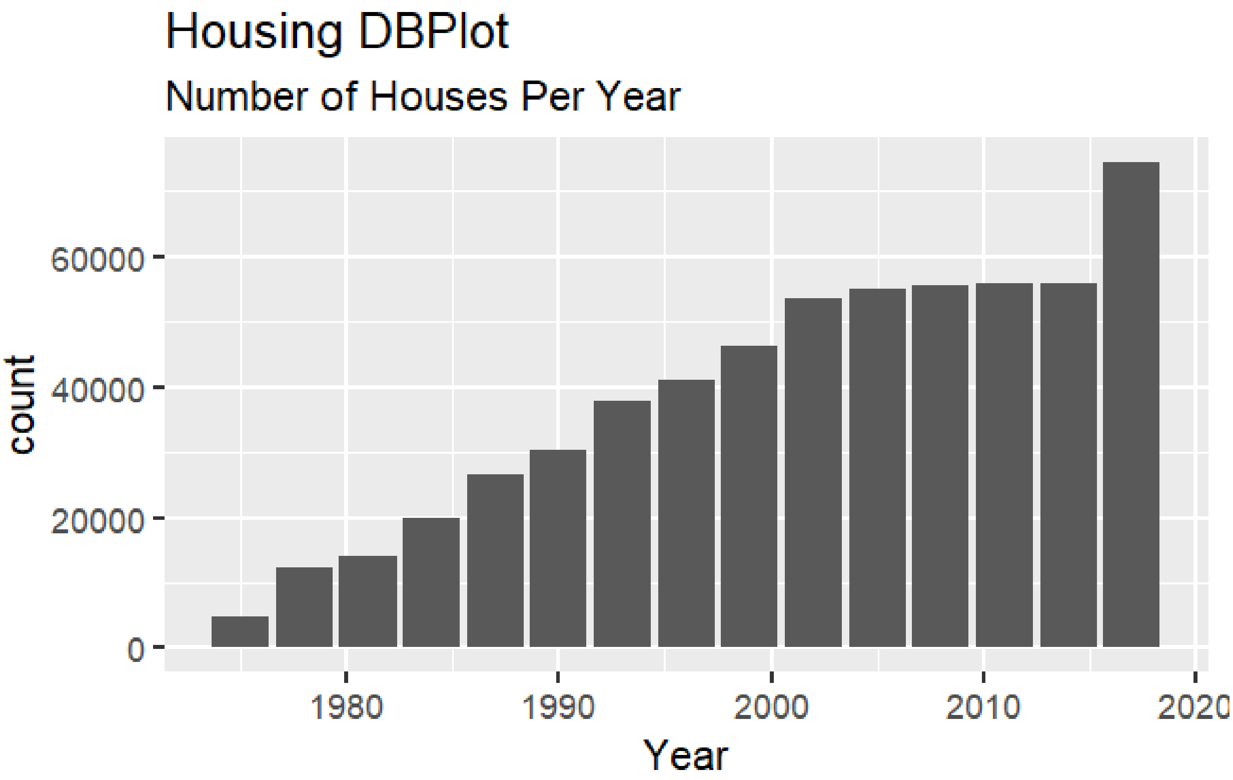
Appendix B



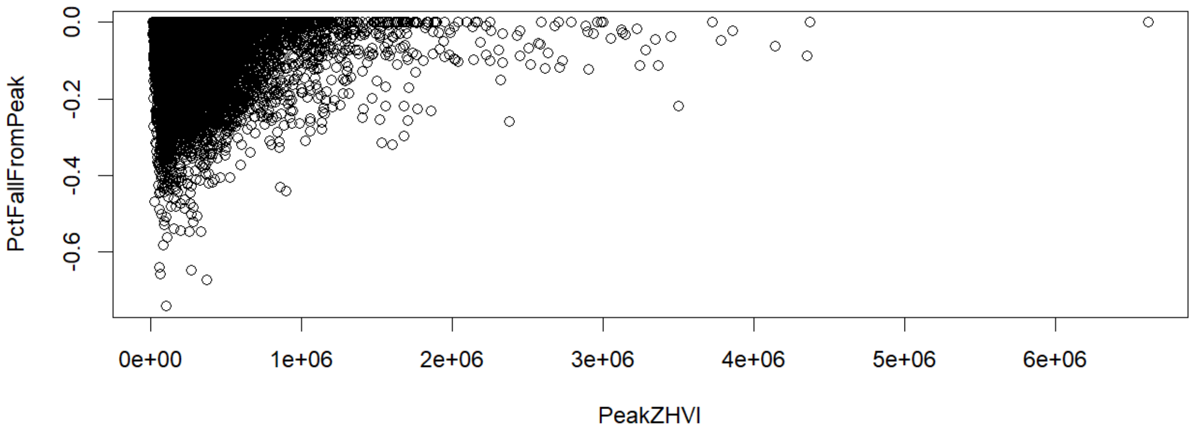
Appendix C



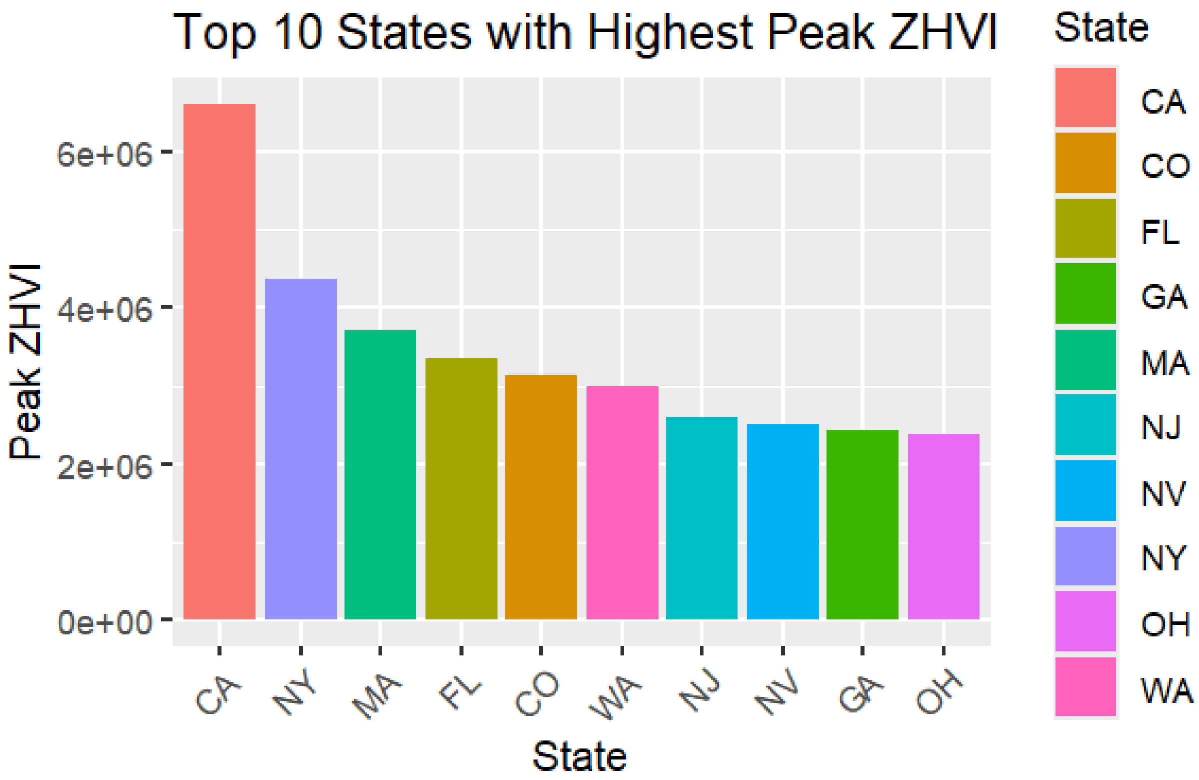
Appendix D



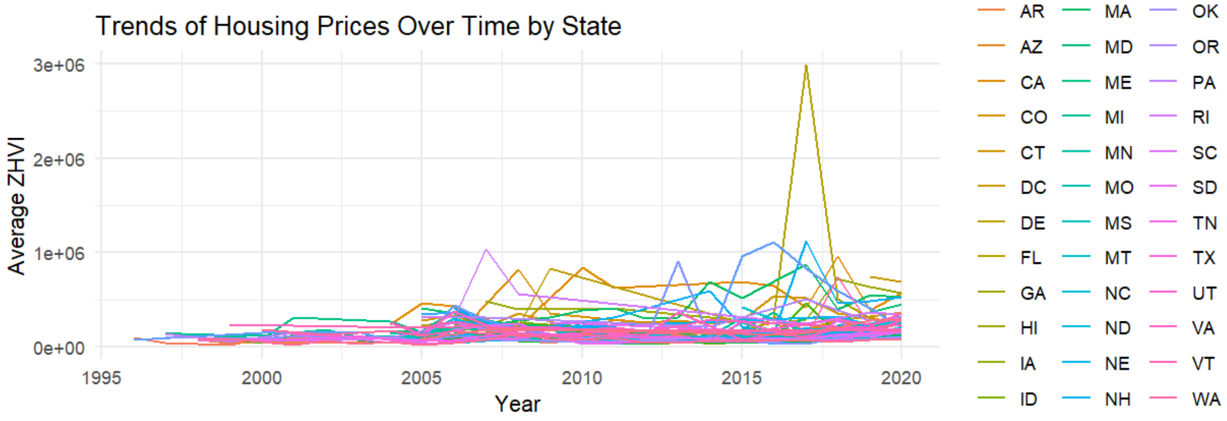
Appendix E



Appendix F



Appendix G


Trends of Housing Prices Over Time by State
Average ZHVI
3e+06
2e+06
1e+06
0e+00
1995
2000
2005
2010
2015
2020
Year
AR
AZ
CA
CO
CT
DC
DE
FL
GA
HI
IA
ID
MA
MD
ME
MI
MN
MO
MS
MT
NC
ND
NE
NH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VA
VT
WA