Name:

# EDA with Zillow Data

For this exploration, you will use data from the real-estate website, Zillow, using Rstudio. You will apply the skills you've learned this year to start answering questions about houses listed on the site using the dataset.

***This document has all of your directions and is where you will record answers. You will use RStudio to calculate statistics and produce visualizations.***

## Part 1: Familiarizing Yourself With the Data

## Background Information about Zillow

The website Zillow estimates the home prices for over 100,000,000 homes around the United States. (Well, actually they call them Zestimates.) In their own words,"We use proprietary automated valuation models that apply advanced algorithms to analyze our data to identify relationships within a specific geographic area, between this home-related data and actual sales prices. Home characteristics, such as square footage, location or the number of bathrooms, are given different weights according to their influence on home sale prices in each specific geography over a specific period of time, resulting in a set of valuation rules, or models that are applied to generate each home's Zestimate. Specifically, some of the data we use in this algorithm include:

Physical attributes: Location, lot size, square footage, number of bedrooms and bathrooms and many other details.

Tax assessments: Property tax information, actual property taxes paid, exceptions to tax assessments and other information provided in the tax assessors' records.

Prior and current transactions: Actual sale prices over time of the home itself and comparable recent sales of nearby homes

Currently, we have data on 110 million homes and Zestimates and Rent Zestimates on approximately 100 million U.S. homes. (Source: Zillow Internal, March 2013) " ———[1]

[1] This write up about Zillow, the data used, and the idea for this project came from the following website: http://community.amstat.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=288c3e05-1ba5-450d-8ec8-62629b876557&forceDialog=0

**Part 2: Setup - Downloading the dataset and the RStudio Worksheet**

1. **Dataset:**

Click the following link, Zillow Housing Dataset, and download the dataset. If you are working on paper, go to Google Classroom and find the dataset in the assignment titled Zillow Data Exploration. Download the dataset, but do not try to open it.

2. **RStudio Worksheet:**

Click on the following link Zillow Exploratory Data Exploration Worksheet, download the file, but do not try and open it. If you are working on paper, go to Google Classroom and find the file in the assignment titled Zillow Data Exploration. Download the dataset, but do not try to open it.

3. **Open RStudio on your computer and when in RStudio**
   - Click Import Dataset, From CSV, Browse, find then the dataset file you saved, **change the name** to something short, click import.
   - Click File, Open File, and then find the RStudio Worksheet you downloaded.
   - Once your Rstudio is set up, return to this document for further details

4. **Run the code in the following section.**

```
#Zillow Exploratory Data Analysis Worksheet
```

```
```{r load the tidyverse library}
library(tidyverse)
library(skimr)
```
```

**Part 3: What is included in the dataset?**

We will narrow our focus to 1000 homes in Saratoga County, New York. The data was collected roughly 10 years ago by students at Williams College.

Let's first figure out where Saratoga County is. Find Saratoga Country on a map by clicking this link - Saratoga County . If you are working on paper, Google Saratoga County.

1. What city is it near?

Now let's familiarize yourself with the dataset.

2. Refer to the following section and change NAME_OF_DATASET to the name you gave your dataset when you loaded it into Rstudio:

```
# Exploring the Data

Use this code to help answer questions in the Exploring the Data section of your worksheet

```{r viewing the data}
#after you run this code be sure to click the arrow to see all variables
head(NAME_OF_DATASET)

```
```

The code in that section will allow you to see the first six rows of your dataset.

3. What are the observational units?

4. How many variables are listed?

5. Write down 4 quantitative variables.

6. Write down 4 categorical variables.

7. Write two questions you could look to answer using this dataset?

**Part 4: Visualize the data**

A good way to start your exploratory data analysis is by making visualizations. We will start by examining the price variable and the area of living variable by making two histograms.

**Price:**

1. Go to the section of the code with the label # Visualize the Data

2. In the following code change NAME_OF_DATASET, QUANTITATIVE VARIABLE, and NUMBER in order to make a histogram with a binwidth of 50000 showing the distribution of prices.

```
### Create a histogram showing the distribution of Price
```

```
```{r Price histogram}
p_hist <- NAME_OF_DATASET %>%
  ggplot(aes(QUANTITATIVE_VARIABLE)) +
  geom_histogram(binwidth = NUMBER, fill = "navy", color = "black") +
  labs(title = "Distribution of House Prices", x = "X-Axis Title", y = "Y-Axis Title") +
  scale_x_continuous(limits = c(0, 800000),
                     breaks = seq(0, 800000, 50000))
p_hist
```
```

3. Change the x-axis title, and y-axis title to fit what is being shown.

4. What does the code “limits = c(0,800000) do to your visualization?

5. What is the shape of the distribution?

6. What price do houses in this dataset tend to be? (This can be a rough estimate)

**Living Area:**

7. Go to the section of the code with the label # Visualize the Data

8. In the following code change NAME_OF_DATASET, QUANTITATIVE VARIABLE, and NUMBER in order to make a histogram with a binwidth of 200 showing the distribution of prices.

```
## Create a histogram showing the distribution of Living Area
```

```
```{r Living Area Histogram}
la_hist <- NAME_OF_DATASET %>%
          ggplot(aes(QUANTITATIVE_VARIABLE)) +
  geom_histogram(binwidth = NUMBER, fill = "red", color = "black") +
  labs(title =" Distribution of Living Areas", x = "X-Axis Title",
       y = "y-Axis Title") +
  scale_x_continuous(limits = c(0, 5600),
                     breaks = seq(0, 5600, 200)) +
  scale_y_continuous(limits = c(0, 300),
                     breaks = seq(0, 300, 50))
la_hist
```
```

9. Why did we change the binwidth from 50,000 to 200?

10. What is the shape of the distribution?

11. What size living area do homes in this dataset tend to have? (This can be a rough estimate)

**Part 5: Summary Statistics to Support Your Visualization**

After making visualizations, you need to generate statistics to enhance your analysis.

Go to the section # Summary statistics for histograms

1. **Let's start by getting summary statistics for price.**

    a. Change NAME_OF_DATASET and QUANTITATIVE_VARIABLE in the following code:

```
##Summary Statistics for Price
```{r Price Summary Stats}

ssprice <- NAME_OF_DATASET %>%
          summarize(avg = mean(QUANTITATIVE_VARIABLE), med = median(QUANTITATIVE_VARIABLE),
                    standard_dev = sd(QUANTITATIVE_VARIABLE),
                    iqr = IQR(QUANTITATIVE_VARIABLE))
ssprice
```
```

    b. Write down each summary statistic that was calculated.

    Mean: Standard deviation:

    Median: Interquartile Range (IQR):

    c. Do these numbers support your earlier claims about where prices tend to be? Explain.

    d. Do these numbers support your earlier claims about the shape of the distribution of prices? Explain.

2. **Now let's look at the summary statistics for living area.**

    a. Change NAME_OF_DATASET and QUANTITATIVE_VARIABLE in the following code:

```
##Summary Statistics for Living Area
```{r Living Area Summary Statistics}
sslarea <- NAME_OF_DATASET %>%
          summarize(avg = mean(QUANTITATIVE_VARIABLE), med = median(QUANTITATIVE_VARIABLE),
                    standard_dev = sd(QUANTITATIVE_VARIABLE,
                                      five_number = fivenum(QUANTITATIVE_VARIABLE)),
                    iqr = IQR(QUANTITATIVE_VARIABLE))
sslarea
```
```

    b. Write down each summary statistics that was calculated.

    Mean: Standard deviation:

    Median: Interquartile Range (IQR):

    c. Do these numbers support your claims earlier about where prices tend to be? Explain.

    d. Do these numbers support your claims earlier about the shape of the distribution of living area? Explain.

    e. Change NAME_OF_DATASET, QUANTITATIVE_VARIABLE1, QUANTITATIVE_VARIABLE2, in order to find even more summary statistics for just the Price and Living.Area variables.

```
##Summary Statistics for Price and Living Area
```{r price and living area summary stats}
ss_area_price <- NAME_OF_DATASET %>%
                 select(QUANTITATIVE_VARIABLE1, QUANTITATIVE_VARIABLE2) %>%
                 skim()
ss_area_price
```
```

    f. Subtract the number under p25 from the number under p75 for each variable.

    Price: p75 - p25 = Living Area: p75 - p25 =

    Are these numbers the same as anything you've calculated already? If so, what?

    g. What percent of houses in Saratoga are less than $259,000?

**Part 6: For the next of this exploration we will examine the relationship between the price of houses and the amount of living space.**

**Visualize the Relationship with a Scatter Plot**

a. Before producing a scatter plot, comment on what you believe the relationship between price and living area will be. Be sure to comment on the strength, direction, and form of the relationship.

   Strength: Direction: Form:

   Summary of all 3 in plain english:

b. When examining the relationship between the price of a house and the living area of a house, which will be the independent and which will be the dependent variable.

   Independent Variable:_____

   Dependent Variable:_____

c. Now make a scatter plot of Price vs. Living Area by changing: NAME_OF_DATASET, IndependentVariable, DependentVariable, Main Title, x-axis title and y-axis title.

```
##Create a Scatter Plot

Now let's see if there is an association between the price and living area. Create a scatter
plot of Price vs Living Area.

```{r scatter plot price vs living area}

pvsa <- NAME_OF_DATASET %>%
          ggplot(aes(x= IndependentVariable, y = DependentVariable)) +
          geom_point(alpha = 1) +
          stat_smooth(method = 'lm') +
          labs(title ="Main Title", x = "Independent Variable", y = "Dependent Variable") +
          scale_y_continuous(limits = c(0, 800000),
                     breaks = seq(0, 800000, 250000))

```
```

d. The plot looks a bit cluttered, change alpha = 1 to alpha = 0.25 and run the code.

e. Were your hypotheses about the strength, direction, and form correct?

**Make a linear model to more specifically explain the relationship between price and living area.**

a. Change the code you think needs to be changed in the following section in order to create the linear model and calculate the correlation coefficient.

```
## Creating the Model

Let's now build a linear model to gain more insight into the relationship between Prices and
Living Area.
```

```
```{r create the model}


NAME_YOUR_MODEL <- lm(DependentVariable~IndependentVariable, NAME_OF_DATASET)
NAME_YOUR_MODEL
cor(NAME_OF_DATASET$DependentVariable, NAME_OF_DATASET$IndependentVariable)

```
```

b. After running that code, the following will appear (without the colors)

- The number below Intercept (in the place where the blue box is in my picture), is the y-intercept of your linear model.
- The number next to Living.Area, (in the place where the yellow box is in my picture) is the slope of your linear model.
- The number below your cor code (in the place where the green box is in my picture), is the correlation coefficient.

```
Call:
lm(formula = Price ~ Living.Area, data = zillow)

Coefficients:
(Intercept)  Living.Area
                          

[1]
```

c. Write the equation for the least squares regression line. (Be sure to write it using proper notation)

d. On average, how much does the price increase for every 1 additional square foot of living area?

e. On average, how much does the price increase for every 100 additional square feet of living area?

f. Which of the numbers from question d and e would be a better number to include in a paper describing the relationship? Explain.

**Part 7: Marginal and Conditional Distributions**

Throughout this exploration, you have examined quantitative variables. Let's now explore a couple of the categorical variables in the dataset: the type of heat and whether or not it is new construction. More specifically, we'll compare the heat sources for newly constructed and not newly constructed houses.

**Make a Visualization**

a. Start by producing a segmented bar graph by changing the code in the following section: The x-axis should represent yes or no for new construction (New.Construct) and the bars should be filled with the type of heat (Heat.Type).

```
##Create a Segmented Bar Graph
```{r stacked bar and two way table}
heat_construction <-  NAME_OF_DATASET %>%
                    ggplot(aes(x = CATEGORICAL_VARIABLE, fill = CATEGORICAL_VARIABLE)) +
                    geom_bar(position = "fill")
heat_construction
```
```

b. What is the major difference that you see?

**Generate Statistics to enhance the description of your visualization.**

c. Change NAME_OF_DATASET and CATEGORICAL_VARIABLE in the code below to make a two-way table.

```
## Create a Two-Way Table
```{r two way table}
heattotals <- NAME_OF_DATASET %>%
            group_by(New.Construct, CATEGORICAL_VARIABLE) %>%
            summarize(num = n()) %>%
            spread(New.Construct, num)
heattotals
```
```

d. How many newly constructed homes are in the dataset?

e. How many houses that are not newly constructed are in the dataset.

f. What proportion of newly constructed homes have ‘Hot Air’ as a type of heat?

g. What proportion of not newly constructed homes have ‘Hot Air’ as a type of heat?

h. What proportion of newly constructed homes have ‘Hot Water as a type of heat?

i. What proportion of not newly constructed homes have ‘Hot Water’ as a type of heat?

**Part 8: Tell a Story**

To conclude this activity, write a clear and concise summary of the findings from your exploratory data analysis.