Lab 03: Exploratory Data Analysis with Statistical Testing

Author: Nigel Nelson

Introduction:

- This lab builds off of *Lab 1: Data Cleaning* by using the resulting cleaned data set as the primary data set for this lab. The data set, *CleanedSacramentorealestatetransactions*, has 919 entries and 14 variables describing real estate transactions in California. This includes the **street** address, the **city**, the **zip**, the **state**, the numbers of **beds**, the number of **baths**, the square footage of residential space(**sq_ft**), the **type** of real estate, the **sale date**, the sale **price**, the **latitude**, the **longitude**, whether it is an **empty_lot**, and the **street_type**. In this lab, Statistical tests are used to analyze the relationships between 3 different combinations of the data set's variables: continuous vs. continuous, continuous vs categorical, and categorical vs. categorical. Through these statistical tests, it is determined which variable pairs have meaningful relationships. This is done so that comparisons can be made with *Lab 2 EDA -- Visualization* and the variable pairs that were determined to be meaningful in that past lab using analysis of the created visualizations. By comparing the results as well as the analysis methods used between these two labs, conclusions can be made about the efficacy of methods used, and in which context the results from the two are meaningful.

Imports: ¶

In [1]:
```
import pandas as pd
import scipy.stats as stats
```

Part I: Review of Statistical Tests

GPA vs. Video Game Playing hypothesis:

- Students who play video games have lower GPA's than students who do not.

Video Game Survey Results:

Plays Video Games	Mean GPA	Std. Dev.	Count
Yes	3.4	1.2	68
No	3.3	1.1	32

Questions:

1. what situations would you use a two-sample t-test? Does the situation describe above meet those criteria? Are there any particular assumptions that the t-test makes that may not hold here?

- A two-sample t-test in used for two samples when there is one measurement variable and one nominal variable, and the nominal value only has two values, as such, this video game example fits these requirements perfectly. This test assumes that the observations in each group are normally distributed and that the data is homoscedastic. For this data set it is possible that neither of these are true. Perhaps the collection method for this survey was not random, and instead the surveys were handed out at a video game club, and at an athletic event. This could result in these assumptions not holding as many athletic teams require students hold a minimum GPA to participate, resulting in a non-normal distribution of GPA, and also result in contrasting standard deviations between the two groups.

2. Null hypothesis:
   - The group that answered they don't play video games will have the same mean GPA as the group who answered that they don’t play video games.
3. Alternative hypothesis:
   - The group that answered they don't play video games will have a different mean GPA than the group who answered that they don’t play video games.

Performing T-test on Video Game Data:

In [2]:
```
p_value = stats.ttest_ind_from_stats(3.4, 1.2, 68, 3.3, 1.1, 32).pvalue
print(f't-test p-value: {p_value}')
```

```
t-test p-value: 0.6908062583072547
```

Interpreting p-value:

- Using a significance threshold of 0.01, the null hypothesis cannot be rejected. This means that the differences in GPAs of the two groups are not statistically significant. This means that for the groups surveyed the hypothesis that those who play video games will have a lower GPA is false. However, this data represents a relatively small sample size and thus more data collection and statistical tests must be conducted in order to definitively reject or accept the null hypothesis of whether video game playing affects GPA.

Part II: Exploring Additional Statistical Tests

Statistical Tests and Their Appropriate Variables:

- Linear Regression:
  - Variables:
    - 2 measurement variables
  - Assumptions:
    - The variables are normally distributed and homoscedastic
- Kruskal-Wallis test:
  - Variables:
    - 1 measurement, 1 nominal variable
  - Assumptions:
    - Observations in each group come from populations with the same shape distribution
- Chi-squared:
  - Variables:

        - 2 nominal variables
    - Assumptions:
        - Individual observations are independent

Statistical Tests and Their Null & Alternative Hypotheses:

- Linear Regression
    - Null Hypothesis:
        - As the X measurement variable gets larger, the Y measurement variable gets neither smaller nor larger.
    - Alternative Hypothesis:
        - The slope of the best fit line between the two measurement variables is not zero.
- Kruskal-Wallis test
    - Null Hypothesis:
        - The mean ranks of the different nominal groups are the same.
    - Alternative Hypothesis:
        - The mean ranks of the different nominal groups are not the same.
- Chi-squared
    - Null Hypothesis:
        - The relative proportions of one of the nominal variables is independent of the second nominal value.
    - Alternative Hypothesis:
        - The relative proportions of one of the nominal variables is dependent of the second nominal value.

Statistical Tests and Implications of Their Null Hypotheses:

- Linear Regression
    - Accepted Null Hypothesis Implication:
        - The value of one of the measurement variables cannot predict the value of the second measurement variable.
    - Rejected Null Hypothesis Implication:
        - The value of one of the measurement variables can predict the value of the second measurement variable to some degree.
- Kruskal-Wallis test
    - Accepted Null Hypothesis Implication:
        - The value of the nominal variable does not determine differences in the measurement variable.
    - Rejected Null Hypothesis Implication:
        - The value of the nominal variable determines differences in the measurement variable to some degree.
- Chi-squared test
    - Accepted Null Hypothesis Implication:
        - The relative proportions of one of the nominal variables is not affected by the second nominal value.
    - Rejected Null Hypothesis Implication:
        - The relative proportions of one of the nominal variables is affected by the second nominal value.

Part III: Regression on Price

Loading Real Estate Transactions:

In [3]:
```
re_transactions = pd.read_csv('CleanedSacramentorealestatetransactions.csv',
                              dtype={'city': 'category', 'zip': 'category',
                                     'state': 'category', 'beds': 'category',
                                     'baths': 'category', 'type': 'category',
                                     'street_type': 'category'})
re_transactions.head()
```

Out[3]:

	street	city	zip	state	beds	baths	sq__ft	type	sale_date	price
0	3526 HIGH ST	SACRAMENTO	95838	CA	2	1	836	Residential	Wed May 21 00:00:00 EDT 2008	59222
1	51 OMAHA CT	SACRAMENTO	95823	CA	3	1	1167	Residential	Wed May 21 00:00:00 EDT 2008	68212
2	2796 BRANCH ST	SACRAMENTO	95815	CA	2	1	796	Residential	Wed May 21 00:00:00 EDT 2008	68880
3	2805 JANETTE WAY	SACRAMENTO	95815	CA	2	1	852	Residential	Wed May 21 00:00:00 EDT 2008	69307
4	6001 MCMAHON DR	SACRAMENTO	95824	CA	2	1	797	Residential	Wed May 21 00:00:00 EDT 2008	81900

Fitting Linear Regression Model for Price vs. each Continous Variable:

In [4]:

```
continous_vars = ['sq__ft', 'latitude', 'longitude']

for var in continous_vars:
    slope, intercept, r, p, stderr = stats.linregress(
        re_transactions['price'], re_transactions[var])
    print(f'Linear Regression for Price vs {var}:')
    print(f'Slope: {slope}')
    print(f'Intercept: {intercept}')
    print(f'R-value: {r}')
    print(f'P-value: {p}')
    print(f'Standard Error: {stderr}\n')
```

```
Linear Regression for Price vs sq__ft:
Slope: 0.003273130686612822
Intercept: 668.6845560376739
R-value: 0.5281109752355064
P-value: 3.866350017864997e-67
Standard Error: 0.00017380043959897243

Linear Regression for Price vs latitude:
Slope: -1.252500955040025e-07
Intercept: 38.62485930765487
R-value: -0.11418187627856692
P-value: 0.0005242263615858759
Standard Error: 3.598707494737652e-08

Linear Regression for Price vs longitude:
Slope: 1.927639930411026e-07
Intercept: -121.41374242428904
R-value: 0.20210937828889763
P-value: 6.308608183945195e-10
Standard Error: 3.084597149477987e-08
```

Table for Linear Regression Models of Price vs. each Continous Variable:

Variable Name	P-value	Statistically Significant ($\alpha = 0.01$)
Square Feet	3.8664e-67	Yes
Latitude	0.0005	Yes
Longitude	6.3086e-10	Yes

Using Kruskal-Wallis test for Price vs. Categorical Variables:

In [5]:
```
categorical_vars = ['city', 'zip', 'beds', 'baths', 'type', 'empty_lot', 'str
for var in categorical_vars:
    samples_by_group = []
    for value in set(re_transactions[var]):
        mask = re_transactions[var] == value
        samples_by_group.append(re_transactions["price"][mask])
    stat, p = stats.kruskal(*samples_by_group)
    print(f'Kruskal-Wallis test for Price vs. {var}:')
    print(f'Statistic: {stat}')
    print(f'P-value: {p}\n')
```

```
Kruskal-Wallis test for Price vs. city:
Statistic: 301.5262572311318
P-value: 3.3269431060533453e-43

Kruskal-Wallis test for Price vs. zip:
Statistic: 467.28164098275016
P-value: 9.291819037186042e-62

Kruskal-Wallis test for Price vs. beds:
Statistic: 168.879592563933
P-value: 4.324627349801965e-33

Kruskal-Wallis test for Price vs. baths:
Statistic: 208.17001369562053
P-value: 5.0716060872070855e-43

Kruskal-Wallis test for Price vs. type:
Statistic: 27.74598075382496
P-value: 4.106383458904976e-06

Kruskal-Wallis test for Price vs. empty_lot:
Statistic: 7.329317423439557
P-value: 0.006783881144183107

Kruskal-Wallis test for Price vs. street_type:
Statistic: 100.1436855081538
P-value: 8.367606094247467e-14
```

Table of Kruskal-Wallis test results for Price vs. Categorical Variables:

Variable Name	P-value	Statistically Significant (α=0.01)
City	3.3269e-43	Yes
Zip	9.2918e-62	Yes
State	N/A	No
Beds	4.3246e-33	Yes
Baths	5.0716e-43	Yes
Type	4.1064e-6	Yes
Empty_lot	0.0068	Yes

Variable Name	P-value	Statistically Significant (α=0.01)
Street_type	8.3676e-14	Yes

Comparing Statistical Results with Lab 2's Graphical Analysis Method:

- *Variables Determined to be Predictive of Price in Lab 2*:
  - City
  - Zip
  - Type
  - Square Feet
  - Beds
  - Baths
- *Variables Determined to be Predictive of Price using Statistical Tests*:
  - Square Feet
  - **Latitude**
  - **Longitude**
  - Zip
  - City
  - Beds
  - Baths
  - Type
  - **Empty_lot**
  - **Street_type**
- Overall, there is a significant amount of overlap between the variables determined to be predictive of price by both the graphical analysis and statistical testing methods. However, the graphical analysis method did not determine that latitude, longitude, empty_lot, and street_type were predictive of price while the statistical testing method did. These differences are likely due to the binary nature of the statistical test's hypotheses in comparison to the more nuanced evaluation used in lab 2. Latitude and Longitude were determined to be significant predictors of price by using linear regression, in which the null hypothesis states that the line of best fit between the two variables is 0. So, no matter how loose the line of best fit is, as long as it's slope is not 0 with some degree of certainty the null hypothesis can be rejected. So while latitude and longitude do not have a visually clear linear relationship with price, the underlying line of best fit is not 0 to a reasonable degree of certainty. As for empty_lot and street_type, the statistical test used in the Kruskal-Wallis test, whose null hypothesis in this case is that the average rank of prices for each of the groups contained in the categorical variable is the same. So in order to reject this hypothesis, the average price rank of each variable's contained groups just had to be marginally different, even if it isn't obvious through visual analysis.

Part IV: Classification on Property Type:

Running Kruskal-Wallis test for Property Type vs. each Continous Variable:

In [6]:
```
continous_vars = ['sq__ft', 'price', 'latitude', 'longitude']
for var in continous_vars:
    samples_by_group = []
    for value in set(re_transactions['type']):
        mask = re_transactions['type'] == value
        samples_by_group.append(re_transactions[var][mask])
    stat, p = stats.kruskal(*samples_by_group)
    print(f'Kruskal-Wallis test for Type vs. {var}:')
    print(f'Statistic: {stat}')
    print(f'P-value: {p}\n')
```

```
Kruskal-Wallis test for Type vs. sq__ft:
Statistic: 62.24776577746572
P-value: 1.9448556846070265e-13

Kruskal-Wallis test for Type vs. price:
Statistic: 27.74598075382496
P-value: 4.106383458904976e-06

Kruskal-Wallis test for Type vs. latitude:
Statistic: 2.149429890370203
P-value: 0.5419776629168942

Kruskal-Wallis test for Type vs. longitude:
Statistic: 4.100386894960272
P-value: 0.25082620468236794
```

Table of Kruskal-Wallis test results for Type vs. Continous Variables:

Variable Name	P-value	Statistically Significant (α=0.01)
Square Feet	1.9449e-13	Yes
Price	4.1063e-6	Yes
Latitude	0.5420	No
Longitude	0.2508	No

Running Chi-squared test for Type vs. Categorical Variables:

In [7]:
```
categorical_vars = ['city', 'state', 'zip', 'beds', 'baths', 'empty_lot', 'st
for var in categorical_vars:
    combination_counts = re_transactions.value_counts(subset=["type", var]
                                                       ).unstack(level=0).fillr
    chi2, p, _, _ = stats.chi2_contingency(combination_counts)
    print(f'Chi-squared test for Type vs. {var}:')
    print(f'Test Statistic: {chi2}')
    print(f'P-value: {p}\n')
```

```
Chi-squared test for Type vs. city:
Test Statistic: 979.9832497484371
P-value: 4.297296313399103e-139

Chi-squared test for Type vs. state:
Test Statistic: 0.0
P-value: 1.0

Chi-squared test for Type vs. zip:
Test Statistic: 511.9447214422208
P-value: 1.172498738482317e-29

Chi-squared test for Type vs. beds:
Test Statistic: 367.4518192362041
P-value: 4.864709422019391e-65

Chi-squared test for Type vs. baths:
Test Statistic: 243.2693224098068
P-value: 3.013838326163985e-43

Chi-squared test for Type vs. empty_lot:
Test Statistic: 10.135168912828723
P-value: 0.017451412991002547

Chi-squared test for Type vs. street_type:
Test Statistic: 188.93785406511302
P-value: 1.0139956378903576e-17
```

Table of Chi-squared test results for Type vs. Categorical Variables:

Variable Name	P-value	Statistically Significant (α=0.01)
City	4.2973e-139	Yes
State	1.0	No
Zip	1.1725e-29	Yes
Beds	4.8647e-65	Yes
Baths	3.0138e-43	Yes
Empty_lot	0.0175	No
Street_type	1.0140e-17	Yes

Comparing Statistical Results with Lab 2's Graphical Analysis Method:

- *Variables Determined to be Predictive of Type in Lab 2*:
  - Price
  - Square Feet
  - Beds
  - Baths
- *Variables Determined to be Predictive of Type using Statistical Tests*:
  - Price
  - Square Feet
  - **City**
  - **Zip**
  - Beds
  - Baths
  - **Street_type**
- Overall, there is a significant amount of overlap between the variables determined to be predictive of type by both the graphical analysis and statistical testing methods. However, the graphical analysis method did not determine that city, zip, and street_type were predictive of price while the statistical testing method did. These differences are likely due to the binary nature of the statistical test's hypotheses in comparison to the more nuanced evaluation used in lab 2. City, zip, and street_type were determined to be predictive of type by using the Chi-squared test, whose null hypothesis is that proportions of one of the variables is independent of the second variable's value. So, in order to reject this hypothesis, one of these categorical variables simply had to be slightly dependent on the second categorical variables, even if this relationship isn't visually obvious.

Conclusion:

- This lab acted as an exercise in using statistical tests in order to determine meaningful relationships between a data set's variables. In order to account for the all of the different combinations of variable types that could have relationships, the following statistical tests were used to determine inter-variable relationships: Linear Regression, the Kruskal-Wallis test, and the Chi-squared test. From these tests, it was determined that for **price**, the variables square feet, latitude, longitude, zip, city, beds, baths, type, empty_lot, street_type were all predictive to some degree. As for **type**, it was determined that price, sq_ft, city, zip, beds, baths, and street_type were all predictive to some degree. Compared to Lab 2's results, these conclusions were mostly consistent. However, the statistical test concluded that several variables were predictive that were not considered predictive in the results of Lab 2. The likely reason for this difference is that the statistical tests are binary in nature, either the variables have no correlation at all, or there is some correlation, even if it is a very minute amount. In contrast, Lab 2's analysis method was much more nuanced as visual inspection and ones own opinion was used to determine if variables were predictive of each other.