Lab 03: Exploratory Data Analysis (EDA) with Statistical Testing
CS3300 Data Science

Learning Outcomes

1. Understand the basic process of data science and exploratory data analysis including modes of inquiry (hypothesis driven, data driven, and methods driven).

2. Select, apply, and interpret appropriate visual and statistical methods to analyze distributions of individual variables and relationships between pairs of variables.

3. Communicate findings through generated data visualizations and reports.

4. Identify correlated and predictive variables.

Overview

In the last two labs, you cleaned a data set of real estate transactions and then using visualizations to explore the relationships between the independent (input) and dependent (output) variables for two machine learning tasks. In this lab, you are going to test for correlation and association between the independent and dependent variables using statistical tests.

You should prepare your results as a Jupyter notebook. In addition to code, you should have text offering interpretations and explanations. Your notebook should be organized into sections with appropriate headers. The notebook and its code should be clean and polished. Use the Blood Glucose Tutorial as a template and reference.

Instructions

Part I: Review of Statistical Tests

You were introduced to statistical hypothesis testing in your probability and statistics class. We're going to start by reviewing statistical testing.

Let's say that you decide you want to know if playing video games impacts students' grades. You set up a survey which asks students two questions:

1. Do you play video games regularly? Yes / no
2. What is your GPA?

a. Before looking at the survey results, you decide to create a hypothesis. Write your hypothesis in your notebook (one sentence).

You now decide to look at the survey results. You have 100 responses! 68 students said they play video games regularly, while 32 students said they did not. The 68 games have an average GPA of 3.4 with a standard deviation of 1.2, while the 32 non-gamers have an average GPA of 3.3 with a standard deviation of 1.1.

Read the entry for two sample *t*-tests in the Handbook of Biological Statistics and answer the following questions.

b. In 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?

c. Write down the null and alternative hypotheses of the *t*-test for your problem (one sentence each).

d. Use the ttest_ind_from_stats function in Scipy to perform a *t*-test on your data above and report the *p*-value. Interpret your *p*-value using a significance threshold (α) of 0.01. Are you able to reject the null hypothesis? Are the differences in GPAs of the two groups statistically significant?

e. How does the evidence from the statistical test relate to your hypothesis in (a)?

Hypotheses for statistical tests are written such that they mutually exclusive and one of the two hypotheses must be true. If the *p*-value from the *t*-test is small enough to reject the null hypothesis, you would accept the alternative hypothesis. It should also be noted that the *t*-test only tells you if the distributions are different; it does not tell you which distribution has a larger mean. If you accepted the alternative hypothesis, you would do a post hoc analysis to determine how the two distributions are different.

Part II: Exploring Additional Statistical Tests

In this lab, we are going to use three statistical tests: a test for correlation based on linear regression, the Kruskal-Wallis test, and a Chi-squared test of goodness of fit. Read the associated entry for each in the Handbook for Biological Statistics.

For each test:

a. List the two types of variables for which the test is appropriate. Indicate any assumptions that you would need to be aware of.

b. Write down the general forms of the null and alternative hypotheses (one sentence per hypothesis).

c. In your own words, write what it would mean if the test did and did not indicate statistical significance.

Part III: Regression on Price

In part III, you are going to explore which variables are predictive of the price (serving as dependent variable).

a. Load the CSV file of the cleaned data set you created in Lab 1.

b. For each continuous variable, use the scipy.stats.linregress() to fit a simple (one variable) linear regression model, estimate the Pearson's correlation coefficient ®, and the statistical significance (*p*-value) of the correlation against the price of the property.

```
slope, intercept, r, p, stderr =
stats.linregress(df["price"], df["latitude"])
```

In a table, indicate the variable name, *p*-value, and whether there is a statistically significant relationship between that variable and price at a threshold of $\alpha = 0.01$.

c. We can test for association between categorical and continuous variables using a Kruskal-Wallis test using the Scipy kruskal() function. In this example, we want to know if the distribution of prices for condos is different from the distribution for other property types:

i. Use Pandas masks to select the prices for each type of property

```
samples_by_group = []
for value in set(df["type"]):
     mask = df["type"] == value
     samples_by_group.append(df["price"][mask])
```

ii. Perform Kruskal-Wallis test:

```
stat, p = stats.kruskal(*samples_by_group)
```

In a table, indicate the variable name, *p*-value, and whether there is a statistically significant relationship between that variable and price at a threshold of $\alpha = 0.01$.

d. How do the results of the statistical tests compare with your analysis of the visualizations of variable pairs from Lab 2?

Part IV: Classification on Property Type

In part IV, you are going to explore which variables are predictive of the property type (serving as dependent variable).

a. Run Kruskal-Wallis test for each continuous variable versus the property type. In a table, indicate the variable name, *p*-value, and whether there is a statistically significant relationship between that variable and property type at a threshold of $\alpha = 0.01$.

b. We can test two categorical variables for association using a Chi-squared test of independence. The "normal" Chi-squared goodness-of-fit test tests if one set of categorical counts was generated from the same distribution as a second set of categorical counts. The test can also be used to test for independence of two variables (meaning that they have no relationship). In the second scenario, expected frequencies of co-occurrences of the values from the two variables are calculated under the assumption that the values are independent. The "normal" Chi-squared test is then used to determine if the co-occurrence counts of the other data set match the expected independent distribution (null hypothesis). If the counts do not match, then you reject the null hypothesis and accept that the alternative hypothesis, which states that the two variables are not independent. Scipy implements the Chi-squared test of independence for us (including the calculation of the expected frequencies under the assumption of independence) as the chi2_contingency() function:

```
combination_counts = df[["type","Street Type"]]
                     .groupby(by=["type","Street Type"])
                     .size()
                     .unstack(level=0)
chi2, p, _, _ = stats.chi2_contingency(combination_counts)
```

Run a Chi^2 test of independence between each categorical variable versus the property type. In a table, indicate the variable name, *p*-value, and whether there is a statistically significant relationship between that variable and property type at a threshold of $\alpha = 0.01$.

c. How do the results of the statistical tests compare with your analysis of the visualizations of variable pairs from Lab 2?

Submission Instructions

Save the Jupyter notebook as a PDF and upload that file through Canvas.

Rubric

Followed submission instructions	5%
Part I: Review of Statistical Testing	
Generated hypotheses and comparison to test results	5%
Appropriateness of *t*-test	5%
Statistical Hypotheses	5%
Running and interpreting t-test results	5%
Part II: Exploring Additional Statistical Tests	
Test description (x3)	5%
Test hypotheses (x3)	5%
Interpretation of test (x3)	5%
Part III: Regression	
Correlation tests (with table and correct interpretations of *p*-values)	10%
Kruskal-Wallis tests (with table and correct interpretations of *p*-values)	10%
Comparison of results with Lab 2	5%
Part IV: Classification	
Kruskal-Wallis tests (with table and correct interpretations of *p*-values)	10%
Chi-squared tests (with table and correct interpretations of *p*-values)	10%
Comparison of results with Lab 2	5%
Formatting: Report is polished and clean. No unnecessary code. Section headers are used. Plots are described and interpreted using text. The report contains an introduction and conclusion.	5%
Exceeded Expectations	5%