**Introduction to Applied Data Science**

**Homework 2**

**Q1. Using the data given below, perform a 2-sample t-test as shown at** [**https://www.statology.org/interpret-t-test-results-in-excel/**](https://www.statology.org/interpret-t-test-results-in-excel/)**, or use any other tool you are familiar with using.**

A: 14,15,15,16,15,8,14,17,16,14,19,20,21,15,15,18,16,12,14,12

B: 15,17,14,17,13,9,10,19,19,14,17,22,23,17,13,16,14,18,25,13

|  |  |  |
| --- | --- | --- |
|  | Variable 1 | Variable 2 |
| Mean | 15.3 | 16.25 |
| Variance | 8.536842105 | 16.61842105 |
| Observations | 20 | 20 |
| Pooled Variance | 12.57763158 |  |
| Hypothesized Mean Difference | 0 |  |
| df | 38 |  |
| t Stat | -0.847079498 |  |
| P(T<=t) one-tail | 0.201127195 |  |
| t Critical one-tail | 1.68595446 |  |
| P(T<=t) two-tail | 0.402254389 |  |
| t Critical two-tail | 2.024394164 |  |

**Q2. Using the survey data given below, perform a chi-squared test as shown at** [**https://real-statistics.com/chi-square-and-f-distributions/independence-testing/**](https://real-statistics.com/chi-square-and-f-distributions/independence-testing/)**, or use any other tool you are familiar with using.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **High Salary** | **Medium Salary** | **Low Salary** | **Total** |
| **State A** | **25** | **45** | **10** | **80** |
| **State B** | **5** | **50** | **60** | **115** |
| **State C** | **50** | **30** | **25** | **105** |
| **Total** | **80** | **125** | **95** | **300** |

With State A as Observed and State B as expected the p value is 2.96E-27, leading us to reject the null hypothesis.

**Q3. Using the data given below, perform linear regression and polynomial regression as shown at** [**https://realpython.com/linear-regression-in-python/**](https://realpython.com/linear-regression-in-python/)**, or use any other tool you are familiar in using.**

ture Yield

0 50 3.3

1 50 2.8

2 50 2.9

3 70 2.3

4 70 2.6

5 70 2.1

6 80 2.5

7 80 2.9

8 80 2.4

9 90 3.0

10 90 3.1

11 90 2.8

12 100 3.3

13 100 3.5

14 100 3.0

Linear Regression:

y = 0.007x + 2.306

R^2: 0.0924176456091349

MSE: 0.1327087087087087

Polynomial Regression:

y = -0.154x^2 + 0.001x + 7.96

R^2: 0.6732052768464262

MSE: 0.04778465063001145

**Q4. Given the data in the file numerics.csv, find the statistical data as shown in the slide titled “Problem With Numerics” in Exploratory Data Analysis. Check also what happens when you change one of the data (value = 3) by multiplying it by 2 and then by 20.**

Original (12, 25, 7, 5, 10, 23, 5, 6, 27, 3, 13, 13, 10, 18, 5):

Mean: 12.133333333333333

Median: 10.0

Count: 15

Minimum: 3

Maximum: 27

Range: 24

Standard Deviation (sd): 7.526545614615571

Variance: 56.64888888888889

First Quartile (Q1): 5.5

Third Quartile (Q3): 15.5

Multiplied by 2 (12, 25, 7, 5, 10, 23, 5, 6, 27, 6, 13, 13, 10, 18, 5):

Mean: 12.333333333333334

Median: 10.0

Count: 15

Minimum: 5

Maximum: 27

Range: 22

Standard Deviation (sd): 7.3181661333667165

Variance: 53.55555555555556

First Quartile (Q1): 6.0

Third Quartile (Q3): 15.5

Multiplied by 20 (12, 25, 7, 5, 10, 23, 5, 6, 27, 60, 13, 13, 10, 18, 5):

Mean: 15.933333333333334

Median: 12.0

Count: 15

Minimum: 5

Maximum: 60

Range: 55

Standard Deviation (sd): 13.762105782021717

Variance: 189.39555555555555

First Quartile (Q1): 6.5

Third Quartile (Q3): 20.5

**Q5. Using the same dataset as Question 4, draw a box plot, a pie chart, a line graph, and bar graph.**

data = [12, 25, 7, 5, 10, 23, 5, 6, 27, 3, 13, 13, 10, 18, 5]

# Boxplot

plt.figure(figsize=(6, 4))

plt.boxplot(data)

plt.title('Boxplot')

plt.xlabel('Data')

plt.show()

# Pie Chart

plt.figure(figsize=(6, 6))

plt.pie(data)

plt.title('Pie Chart')

plt.show()

# Line Graph

plt.figure(figsize=(6, 4))

plt.plot(range(1,16), data, marker='o')

plt.title('Line Graph')

plt.xlabel('X-axis')

plt.ylabel('Y-axis')

plt.show()

# Bar Graph

plt.figure(figsize=(8, 4))

plt.bar(range(1,16), data)

plt.title('Bar Graph')

plt.xlabel('Categories')

plt.ylabel('Values')

plt.show()