Updated May 2024

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/, 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

t-Test: Two-Sample Assuming Equal Variances

	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/, or use any other tool you are familiar witgh 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**

Expected Values

	High Salary	Medium Salary	Low Salary	Total
State A	21.33333333	33.33333333	25.33333333	80
State B	30.66666667	47.91666667	36.41666667	115
State C	28	43.75	33.25	105
Total	80	125	95	300

Updated May 2024

Chi-Square Test

SUMMARY		Alpha	0.05
Count	*Rows*	*Cols*	*df*
300	3	3	4

CHI-SQUARE					
	chi-sq	*p-value*	*x-crit*	*sig*	*Cramer V*
Pearson's	74.49334	2.55E-15	9.487729	yes	0.352357
Max likelihood	83.47093	3.2E-17	9.487729	yes	0.372986

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

i	**Temperature**	**Yield**
1	50	3.3
2	50	2.8
3	50	2.9
4	70	2.3
5	70	2.6
6	70	2.1
7	80	2.5
8	80	2.9
9	80	2.4
10	90	3
11	90	3.1
12	90	2.8
13	100	3.3
14	100	3.5
15	100	3

Linear Output:

- **coefficient of determination**: 0.09241764560913446
- **intercept**: 2.306306306306306
- **slope**: [0.00675676]

Polynomial Output:

- **coefficient of determination**: 0.6732052768464252
- **intercept**: 0.0
- **coefficients**: [7.96048110e+00 -1.53711340e-01 1.07560137e-03]

Q4. Given the data set below, find the statistical data as shown in the slide titled “Problem With Numerics” in Exploratory Data Analysis (Week 4, Slide20). Check also what happens when you change one of the data (value = 3) by multiplying it by 2 and then by 20.

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

Original:

- *Mean = 12.13* Median = 10.0 Count = 15
- Min = 3 Max = 27 Range = 24
- *S.D. = 7.53* *Variance = 56.65*
- First quartile = 5.5 Third quartile = 15.5

(value = 3 * 2):

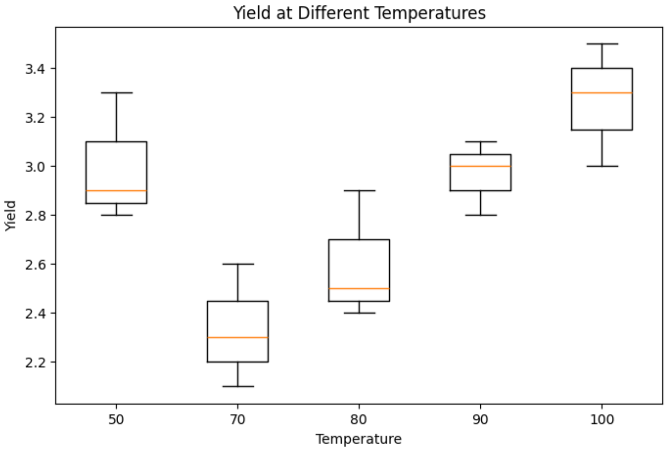
- *Mean = 12.33* Median = 10.0 Count = 15
- Min = 5 Max = 27 Range = 22
- *S.D. = 7.32* *Variance = 53.56*
- First quartile = 6 Third quartile = 15.5

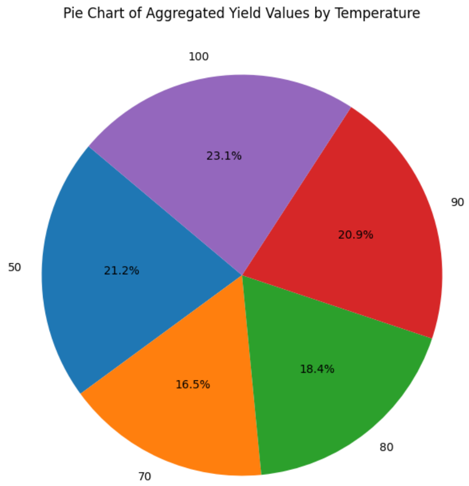
(value = 3 * 20):

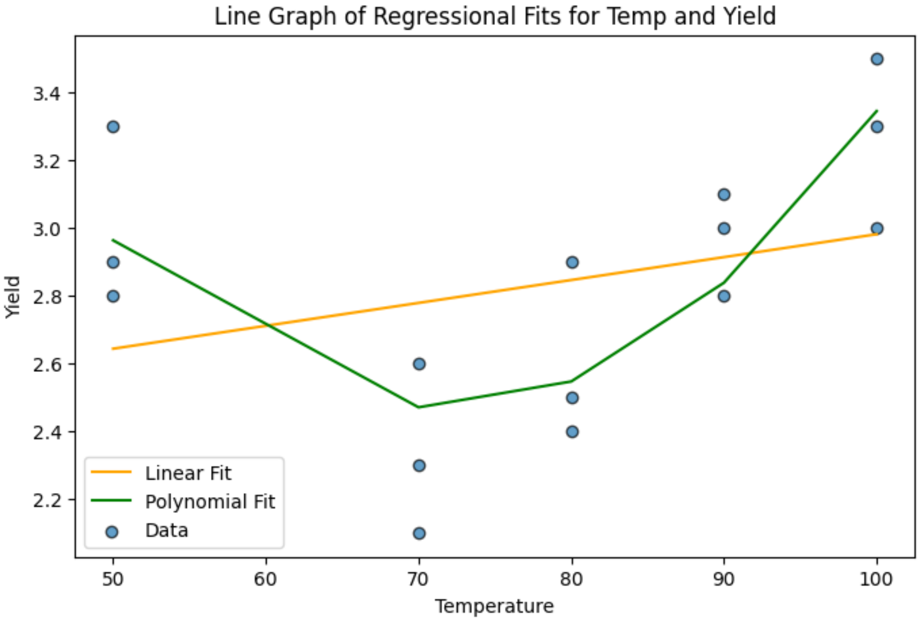
- *Mean = 15.93* Median = 12.0 Count = 15
- Min = 5 Max = 60 Range = 55
- *S.D. = 13.76* *Variance = 189.40*
- First quartile = 6.5 Third quartile = 20.5

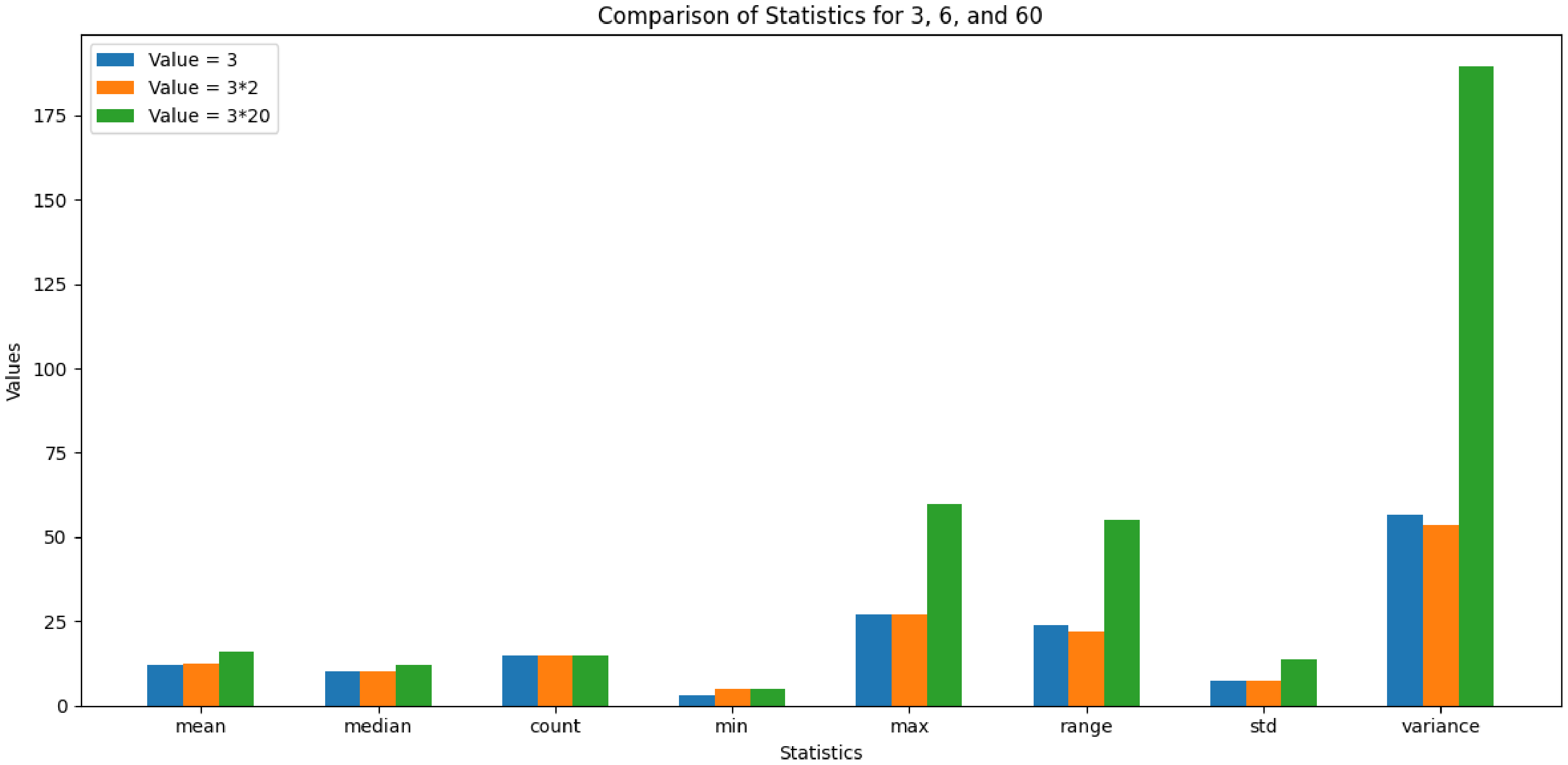
Q5. Using the same datasets in Questions 3 or 4, draw a box plot, a pie chart, a line graph, and bar graph. Label your charts, and identify which data was used to make the chart in your labels.

Updated May 2024


Yield at Different Temperatures


Pie Chart of Aggregated Yield Values by Temperature


Line Graph of Regressional Fits for Temp and Yield


Comparison of Statistics for 3, 6, and 60
Value = 3
Value = 3*2
Value = 3*20
Values
0
25
50
75
100
125
150
175
mean
median
count
min
max
range
std
variance
Statistics