Due: Monday, February 6, 2023, 11:49 p.m.

Data Analytics and Modeling: Quantitative Analysis for Economic Strategy (ECON453)

Spring, 2023

Problem Set 1

Instructions:

- Submit your well labeled R code as an addendum along with your write-up. Start your R code with a comment with "your name, Econ453, pset 1."
- I will look at your handwritten (hopefully typed) answers first. Your handwritten answers should be self contained. I will go over your code to see your technique and to see that your code matches with your answers. Please do not just dump output from R code and expect me to hunt down the answers.
- Where needed, data tables are available on d2L in an EXCEL file.
- Upload your completed homework as a single PDF file to the d2L. Please do not e-mail or submit hard copies.

1. The following table gives standardized test scores for randomly selected fourth graders from two schools (Lincoln School and Kennedy School) and for two years (2014 and 2016). Only a few observations are shown here; complete data are on d2L in pset1.xlsx, sheet "scores."

year	school	score
2014	Lincoln	40
2014	Lincoln	12
.	.	.
.	.	.
2014	Lincoln	59
2014	Kennedy	66
2014	Kennedy	93
.	.	.
.	.	.
2014	Kennedy	73
2016	Lincoln	65
2016	Lincoln	95
.	.	.
.	.	.
2016	Lincoln	82
2016	Kennedy	68
2016	Kennedy	85
.	.	.
.	.	.
2016	Kennedy	99
2016	Kennedy	51

a) Provide summary statistics for scores. Provide min, max, sample mean, sample variance, sample standard deviation, coefficient of variation, mean absolute deviation, Q1, median, Q3, and IQR. Present your results in well formatted table with a title.

b) Draw a Box and whisker plot for score.

c) Calculate summary statistics (this time only sample mean, and sample standard deviation) for scores by year and by school. Present these numbers in well formatted table and a diagram (a bar chart would be good).

d) Based on these numbers, how do the schools compare?

2. The following table gives household level survey data on health care expenses, income, and education for households in the US, Canada, and Mexico. Only a few observations are shown here; complete data are on d2L in pset1.xlsx, sheet "medical_expenses."

COUNTRY	LOCATION	MEDICALEXPN (100$)	INCOME (1000$)	EDUCATION (Years)
USA	RURAL	22.265	19	12
USA	RURAL	8.98	42	10
.	.	.	.	.
USA	URBAN	48.061	64	17
.	.	.	.	.
CANADA	RURAL	2.287	23	5
.	.	.	.	.
CANADA	URBAN	20.722	29	13
CANADA	URBAN	38.404	70	15
.	.	.	.	.
MEXICO	RURAL	13.065	7	3
MEXICO	RURAL	15.457	28	12
.	.	.	.	.
MEXICO	URBAN	49.915	76	18
MEXICO	URBAN	9.104	9	5
.	.	.	.	.

a) Provide summary statistics for all numeric variables. Provide min, max, sample mean, sample variance, sample standard deviation, coefficient of variation, mean absolute deviation, Q1, median, Q3, and IQR. Present your results in well formatted table with a title.

b) Are there any outliers in the data for medical expenses? If so, identify them using suitable R code. Note: you may use one of these criteria to identify outliers (i) data points that are more than three standard deviations away from mean (ii) data point that is outside [Q1-1.5•IQR, Q3+1.5•IQR]. In your answer, indicate which method you used.

c) Calculate summary statistics for all numeric variables (this time only sample mean, and sample standard deviation) by country and by location (urban/rural). Present these numbers in well formatted table and a diagram (a bar chart would be good). You may consider three separate diagrams for the three numeric variables.

d) Based on these numbers, how do the countries compare? How do rural and urban households compare?

e) Draw a scatter plot of medical expenses (on y-axis) and income (on x-axis). What can you infer from this diagram.

f) Calculate sample correlations between all numeric variables and present them in a table. Comment on the magnitude and sign of the correlation.