Fox School of Business
TEMPLE UNIVERSITY®

MIS 0855 Data Science (Section 002) – Spring 2015
Assignment #1 – Creating a Data Analysis Plan
Due by Friday, February 6th, 11:59 PM EST

Task:

Develop a plan for data analysis by forming hypothesis and finding data sets that will allow you to test those hypotheses.

Scenario:

Once Temple students graduate, it's time for them to go get a job. But is staying in the Philadelphia area the best choice? Evaluate Philadelphia as a place to live, work, and play compared to the rest of the United States.

Deliverable:

Your deliverable will have three parts. All three parts will be completed using the Deliverable Worksheet at the end of this document.

Part 1: Develop hypotheses – what will you investigate?

Create five hypotheses (testable statements) that would give you greater insight into the issue. For example, "It rains less in Philadelphia than it does in Cleveland."

Each hypothesis should be testable, falsifiable, and grounded in a rationale/theory. It does not have to be true, nor is it your task to demonstrate it's true.

State each hypothesis and its underlying rationale. Remember, the rationale is the reasoning behind the relationship you describe. In other words, why do you believe that crime would be lower in Center City than in other areas of Philadelphia?

Part 2: Identify data sets – where will you find the evidence?

List at least five real data sets or more that provide data relevant to the issue. Each data set by itself may not have all the data you need, and some of the data in each data set might not be relevant.

You may get your data from any of the following sources (it doesn't have to be one from each). Finding a data source on your own that is not listed below is highly encouraged and will result in a higher grade.

- Any source available through Data.gov (http://www.data.gov)
- The United States Census Bureau (http://www.census.gov)
- Any source available through OpenDataPhilly (http://www.opendataphilly.com)
- The Bureau of Labor Statistics (http://www.bls.gov)
- National Climatic Data Center (http://www.ncdc.noaa.gov)

For each data set, list its name and a direct URL where you found it (not just the site URL). If there isn't a direct URL, provide brief instructions how to find it (1-2 sentences).

Give each data set a number – you'll need that for Part 3.

Part 3: Map data to the hypotheses – how will you test?

List the data that you'd use to test each hypothesis. All data must come from the data sets you found in Part 2. You should list the data field (column) name, a brief description of the data, and the data set it came from.

Submission Instruction

- Complete the Deliverable Worksheet. This will contain your responses for parts 1, 2, and 3 of the assignment.
- Submit your completed worksheet into Blackboard by Friday, Feb. 6th, 11:59PM EST. This deadline is firm, and the instructor will not take any extraneous circumstance into consideration that occurs to you such as a PC malfunction or network outages.
- Late submission is allowed, but there will be 10% penalty per each 12 hours. For example, if you submit in the morning of Feb. 8th, a 30% penalty is imposed on your submission. Therefore, your submission will be graded zero after the noon of Wed, Feb 11th.
- Plagiarism : Blackboard SafeAssign detects plagiarism. Plagiarizing other work in any circumstance will be reported to the University immediately as an academic misconduct.

Grading:

Your work will be evaluated using the following criteria:

Category	4 (A-level)	3 (B-level)	2 (C-level)	1 (D or F-level)
Part 1: Develop hypotheses (40%)	• All hypotheses are testable and falsifiable. • The rationale for all hypotheses are stated very clearly and well-reasoned.	• All hypotheses are testable and falsifiable. • The rationale for all hypotheses are stated clearly and are somewhat well-reasoned.	• Some hypotheses are not testable and/or falsifiable. • The rationale for one or more hypotheses is unclear and/or the logic is flawed.	• Most hypotheses are not testable and/or falsifiable. • The rationale for the hypotheses is missing or incomplete.
Part 2: Identify data sets (30%)	• Each data set provides unique, relevant data. • All data sets are properly identified by name and URL. • Instructions are provided if the URL by itself does not take you directly to the data. • Each data set is assigned a number.	• Each data set is relevant but some do not provide unique data. • All data sets are properly identified by name and URL. • Instructions are provided if the URL by itself does not take you directly to the data. • Each data set is assigned a number.	• Some data sets are not relevant to the problem. • All data sets are properly identified by name and URL. • Instructions are missing even when the URL by itself does not take you directly to the data. • Each data set is assigned a number.	• Most data sets are not relevant to the problem. • Some data sets are not properly identified by name and URL. • Instructions are missing even when the URL by itself does not take you directly to the data. • Data sets are not assigned a number.
Part 3: Map data to hypotheses (30%)	• All hypotheses are listed and two or more pieces of data are identified. • The data is strongly relevant to all hypotheses and would allow for a direct test in all cases. • The data for each hypothesis is part of a data set identified in Part 2.	• All hypotheses are listed and two or more pieces of data are identified. • The data is relevant to all hypotheses but in some cases would not allow for a direct test. • The data for each hypothesis is part of a data set identified in Part 2.	• The data for some hypotheses are incomplete (two or more pieces of data are not identified). • For some hypotheses, the data is not relevant; therefore, tests are not possible. • The data for each hypothesis is part of a data set identified in Part 2.	• The data for most hypotheses are incomplete (two or more pieces of data are not identified). • For most hypotheses, the data is not relevant; therefore, tests are not possible. • The data for each hypothesis is not part of a data set identified in Part 2.