Fox School of Business
TEMPLE UNIVERSITY®

MIS 0855 Data Science (Section 006) – Fall 2017
Assignment #1 – Creating a Data Analysis Plan (10% of the Total Grade)
Due by Friday, September 15th, 11:59 PM EST

Please read all the instructions carefully.

Task

Develop a plan for data analysis by developing hypothesis and finding real-life datasets that would allow you to test those hypotheses.

In this assignment, you'll develop your own hypotheses like the ones you've created in the in-class exercise in Day 4. You can develop any interesting hypotheses in any topic of your choosing – sports, politics, economics, education, and so forth.

As a deliverable, you will fill out the accompanied Worksheet and complete Part 1, 2, and 3.

Part 1: Develop hypotheses – what will you investigate?

Create five hypotheses that would provide knowledge and insights for decision makers. For example, "A Philadelphia neighborhood with more tree experiences fewer violent crimes per resident than other neighborhoods" (from Day 4 in-class exercise). A few examples of hypotheses are available below.

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

State each hypothesis as specific as possible and provide sensible rationale(s) for each one. Remember, a rationale is the reasoning behind the relationship you describe. In other words, why do you believe that crime rates would be lower in neighborhoods with more street trees?

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

List real datasets from at least two different data sources that provide data relevant to the issue.

Several recommended data sources are listed below. Finding a data source on your own that is not listed below is highly encouraged and will result in a higher grade.

- Open government data sources such as Census.gov, Data.gov, OpenDataPhilly.org, Washington DC Open Data (http://opendata.dc.gov), NYC Open Data (https://nycopendata.socrata.com/) or

Socrata (https://opendata.socrata.com/).
- Datasets from Pew Research Center (http://www.pewresearch.org/data/)
- Sports statistics, such as those for Major League Baseball (http://www.seanlahman.com/baseball-archive/statistics) or National Football League (http://nflsavant.com/)
- Election data such as FEC (http://www.fec.gov/pubrec/electionresults.shtml)
- Healthcare datasets from https://data.medicare.gov, https://www.healthdata.gov/, or any other source
- Datasets from World Bank (http://data.worldbank.org/)
- Datasets from FiveThirtyEight (http://fivethirtyeight.com/datalab/)
- Any data that you can find on the Web!

For each data set, list its name and a direct URL where the data can be downloaded (not just the site URL). For instance, Philadelphia Street Tree Inventory is located at https://www.opendataphilly.org/dataset/philadelphia-street-tree-inventory, NOT at https://www.opendataphilly.org/. If a direct URL is not available, provide brief instructions how to find it (1-2 sentences).

Give each data set a number, which you'll need 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 datasets you found in Part 2.

A Few Examples of Hypotheses

- Sport : An NFL quarterback who was drafted in the second round scores more touchdowns in the first three years than one who was drafted in the first round.
- Politics/Education : In a city with higher voter turnouts, high-school students receive higher math scores in statewide standardized tests.
- World : A country with a democratically-elected president or prime minister experiences a faster growth in gross domestic product (GDP) than one with an autocratic leader.
- Health : In a city where there are more primary-care doctors per population, residents show a lower level of cholesterol.
- Real Estate : In a suburban neighborhood where SEPTA regional trains run more frequently, a median home price is higher in other neighborhoods.
- Transportation : There are more Uber drivers per population in a city with higher (or lower) unemployment rates.
- Business : A restaurant with a more number of menu items receives fewer Yelp stars.

Submission Instructions

- Complete a Worksheet that contains your responses for Parts 1, 2, and 3 of the assignment.
- Submit your completed worksheet into Canvas by Friday, Sep. 15th, 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 Sep. 17, a 30% penalty is imposed on your submission. Therefore, your submission will be graded zero after the noon of Wed, Sep 20.

Some Tips

- First, select a topic you have been personally interested in, be it sports, politics, economics, stock markets, governments, entertainment, and so on, and browse data sources for the topic.
- It is not necessary to complete Part 1 before Part 2. You can search for datasets and develop hypotheses at the same time.
- You'll get a higher grade when testing a hypothesis requires integrating two datasets, which makes the hypothesis more interesting. For example, the hypotheses in Day 4 in-class exercise require two data sources (Open Data Philly, Philadelphia Police Department) to test.
- You'll get a higher grade from a hypothesis that is counterintuitive and surprising. For instance, a hypothesis that "a neighborhood with a higher violent crime rate also experiences a higher property crime rate" is not quite surprising but rather obvious.
- Be as specific as possible, so that your hypotheses are testable. For instance, "the neighborhood A is safer than B" is not testable, while "a crime rate in A is lower than in B" is testable. A hypothesis must be testable with quantifiable (numerical) values (e.g. crime rates, unemployment rates, profits, the number of votes, the number of goals, …).

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.