FACULTY OF COMPUTING AND INFORMATICS (FCI)
MULTIMEDIA UNIVERSITY
CYBERJAYA.

TDS2101 - INTRODUCTION TO DATA SCIENCE
SESSION - TT02

Lecturer Name: John See Su Yang

Infectious Disease

PART B

Group Members	**Student ID**
Sarah Batrisyia Binti Ahmad Salman	1181301724
Amiera Nur Hafidzah Binti Muzaffar	1181302678

1. Set the context that you would like to examine i.e. the Question(s) you would like to answer.
What is/are your questions? Typically, you can have several questions which are descriptive in nature and at least ONE predictive question. The question(s) may be formulated based on the problem that you have identified, with good context from a particular industry sector, business or society. Although this has been presented in Part A, you can still refine it in Part B.

- What disease X that has affected the most county in California?
- Which county has the most disease rates because of X disease?
- Does infectious diseases correlate to the number of the disease rates?
- Can infectious disease be decreasing in the future?

2. Identify the dataset(s) that you have used and describe the contents.
What is your dataset about? Where did you get it? What kind of information is in the data?

- Datasets that we choose is "Infectious Disease 2001-2014" from "data.world" because it fits our criteria to do data wrangling for this assignment.
  - https://data.world/health/infectious-disease-2001-2014/workspace/file?filename=rows.csv
  - HealthData.gov
- These data contain counts and rates for Centers for Infectious Diseases-related disease cases among California residents by county, disease, sex, and year spanning 2005-2014 (As of September, 2015).
- Inside this data contain the Disease name, County names, Year, Sex, Count, Population, Rate, C.I lower, and C.I upper

Feel free to make alterations if your earlier proposed data was found to be infeasible. For certain cases, you may even have to refine the question formulated previously if you are unable to identify directly relevant or appropriate datasets.

3. Examine the quality of the data and perform the necessary data cleaning.
What are the data cleaning activities/tasks that you have performed? How were they done and why do it?

- The original data contains the data from year 2001 - 2014, but we shorten it since the data is to large.
- We remove the null value for the Count, CI Lower and CI Upper from the data.

4. Explore the data to understand its descriptive statistics.
What are some graphical plots that can help illustrate the current state of the data and are there any interesting correlations in the data?
Construct scripts that can help reveal answers to the descriptive questions that you have asked (if relevant).

5. Employ ONE of the data mining or predictive modelling techniques (e.g. Decision Tree, Naïve Bayes, Linear Regression, K-means Clustering, Logistic Regression etc.) that may be suitable for making some simple predictions.
Are you able to mine some interesting patterns, or build a model to predict future behaviour based on the data you have obtained?
Note: It is not necessary to use machine learning techniques here if you do not know how. However, there is no restriction if you wish to do so.

6. Use compelling visualizations to support a consistent narrative.
Show with visuals how your Question(s) can be answered.
With visuals, it is also easier to discuss and analyse further, making observations that are useful to the target sector.

7. Mention and discuss some of the challenges or restrictions you had faced in this project.
What are some actionable insights that can be done based on your data analysis?
What is the way forward for further insights to be generated?

Answers to Questions:

1. What disease X that has affected the most county in California?

Disease that has affected the most county in California is Chlamydia.

Disease	
Paralytic Shellfish Poisoning	8
Plague, human	12
Rubella	16
Rabies, human	16
Cyclosporiasis	24
Botulism, Other	24
Cholera	24
Psittacosis	48
Botulism, Foodborne	84
Leptospirosis	96
Ciguatera Fish Poisoning	96
Hantavirus Infection	104
Babesiosis	112
Tularemia	120
Anaplasmosis and Ehrlichiosis	160
Tetanus	184
Trichinosis	208
Relapsing Fever	208
Spotted Fever Rickettsiosis	258
Toxic Shock Syndrome (Non-Streptococcal)	296
Streptococcal Infection (cases in food and dairy workers)	340
Hepatitis E, acute infection	372
Leprosy	538
Mumps	688
Botulism, Wound	694
Measles	800
Q Fever	852
Varicella Hospitalizations	852
Scombroid Fish Poisoning	854
Creutzfeldt-Jakob Disease and other Transmissible Spongiform Encephalopathies	878
Brucellosis	998
Hepatitis C, Acute	1040
Cysticercosis or Taeniasis	1458
Typhus Fever	1740
Hemolytic Uremic Syndrome	1750
Shiga Toxin Positive Feces (without culture confirmation)	1812
Dengue	2114
Hepatitis B, Acute	2172
Typhoid Fever, case	2912
Yersiniosis	2970
Hepatitis A	3072
Lyme Disease	3618
Staphylococcus aureus Infection (cases resulting in death or ICU)	3952
Listeriosis	4324
Malaria	4970

Influenza Death (<65 years of age)	5146
Invasive Meningococcal Disease	5522
E. coli Other STEC (non-O157)	5692
Vibrio Infection (non-Cholera)	5734
Legionellosis	7630
E. coli O157	10770
Cryptosporidiosis	13150
Amebiasis	14646
Shigellosis	54794
Giardiasis	78388
Tuberculosis	98818
Coccidioidomycosis	131130
Pertussis	135262
Early Syphilis	178224
HIV	185068
Salmonellosis	191450
Campylobacteriosis	249728
Gonorrhea	1277928
Chlamydia	6142870

2. Which county has the most disease rates because of X disease?