INTRODUCTION TO DATA SCIENCE (F20)

HW2 US Accident Data

This task exercises your ability to use python to visualize a large dataset and get key statistics including mean, standard deviation, and probability. You need to submit the ipynb file to Moodle.

A data collection of traffic accidents in the US are saved in multiple CSV files. Each CSV file has the following columns:

ID	This is a unique identifier of the accident record.
Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic. The data has accidents of severity 1 to 4.
State	Shows the state in the address field. The data has accidents in at least 3 states and at most all 50 US states.
Timezone	Shows timezone based on the location of the accident. The data has accidents of 'US/Pacific', 'US/Eastern', 'US/Mountain', 'US/Central'.
Temperature(F)	Shows the temperature (in Fahrenheit).
Humidity(%)	Shows the humidity (in percentage).
Pressure(in)	Shows the air pressure (in inches).
Visibility(mi)	Shows visibility (in miles).
Crossing	A POI annotation which indicates the presence of crossing in a nearby location.
Stop	A POI annotation which indicates the presence of stop sign in a nearby location.
Traffic_Signal	A POI annotation which indicates the presence of traffic_signal in a nearby location.
Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.

You are asked to write python code to calculate key statistics and build visualizations. The task asks you to solve 6 problems as listed below. For each question, your code should output visualizations and/or CSVs. Write documentation in your code. Add comments to explain key steps. Create one cell for each question. Use the HW2.ipynb to start.

One test case is provided to you. Note the visualizations and CSVs are NOT the correct answers. They just show what needs to be presented in the visualization and CSV. The input file should only be read once in the notebook file. Visualizations need to be saved to png files named question1.png, question2.png, etc. CSVs should be named question1.csv, question2.csv, etc. Visualization should have proper axis ranges, ticks, labels, legend, and markers. Use proper colors and sizes to make the visualization easy to read.

Your code will be tested with multiple hidden test cases (HTCs). For each HTC video file, your code should generate the corresponding answers in Jupytor Notebook. The HTCs are similar to the given test cases. You can assume that HTC video files have all the above-mentioned columns. The cells in your solution will be executed one after another.

- Question 1:

Calculate the means and standard deviations of temperature, visibility, and wind speed grouped by different time zones. For each measurement, use bar charts to visualize means and standard deviations of Temperature(F), Humidity(%), Pressure(in), and Visibility(mi) values of different time zones. The bar heights are the mean values. Use error bars to show the standard deviations. Save the means and standard deviations of the four fields into question1.csv. Sort the result by the ascending order of Timezone. The output CSV should have the following columns:

Timezone	Temperature(F)_mean	Temperature(F)_std	Humidity(%)_mean	Humidity(%)_std	Pressure(in)_mean	Pressure(in)_std	Visibility(mi)_mean	Visibilit

- Question 2:

Visualize a scatter plot of temperature versus humidity of the top 3 states that have the most accidents. You can assume there are at least 3 states and different states have a different number of cases. Save the state name, accident count, and the correlation between temperature and humidity of each state in question2.csv. Sort the result in the descending order of accident count. The output CSV should have the following columns:

State	Accident	Correlation

- Question 3:

Visualize a clustered bar chart of percentages of accidents that happened at a Crossing, Stop, and Traffic_Signal. Group the data by states. Sort the states alphabetically on the x-axis. Each state has 3 clustered bars showing the percentages of accidents at a Crossing, Stop, and Traffic_Signal. Note not all 50 states have records in the data. Your visualization should be dynamically configured to only show the states that have accidents in the data. Save the percentage at each place in a question3.csv. Sort the result in ascending order of state abbreviations. The output CSV should have the following columns:

State	Crossing	Stop	Traffic_Signal

Explanation: If a state does not have a row in the CSV, do not show it in the viz. But if all rows of a state have False for all three places, you should show the state with no bars in the visualization. It means the state has accidents but no accident happened at any of the three places.

- Question 4:

Visualize the PDF function of the distributions of Temperature(F), Humidity(%), and Visibility(mi) by severity. You can assume each severity level has at least 30 records. Then calculate given an accident happens when Temperate = 90F, Humidity is 60%-70%, and Visibility = 4, what is the probability of this accident to be severity level 4? Save the result to the question4.csv. The output CSV should only have one column “answer” and one value of your result.

- Question 5:

Assume any two accidents have the same probability to be one of the severity levels 1, 2, 3, or 4. And any two accidents have the same probability to happen either in the day or at night. Accident severity level and Sunrise_Sunset are independent of each other. For any data set with 100 accidents, what is the probability that more than 60% of them are severity level 1 or 2 and at the same time more than 70% of them happened in the day? Save the result to the question5.csv. The output CSV should only have one column “answer” and one value of your result.

- Question 6:

Build a Naive Bayes classifier function to predict severity by Temperature(F), Visibility(mi), and Sunrise_Sunset. Use the classifier to predict the data in test.csv. Sort the data by the ascending order of ID. Save the prediction in question6.csv. Your code should finish running within 1 minute. The output CSV should have 2 columns:

ID	prediction

Violation of the following output rules will result in losing points:

· Document your code. Explain the key steps in the comments.

· You should only read the input file once. Do not read the CSV file in every question.

· Visualizations must have proper labels, titles, ticks, ranges, colors, sizes, etc. Make it easy to read.

· Visualization needs to be saved to png files named question1.png, question2.png, etc. CSVs should be saved as question1.csv, question2.csv, etc.

· For the questions which ask to visualize the data in ascending or descending orders, the values in visualization must be presented in the required order.

test_case1.zip October 9 2020, 11:32 PM

Submission status

Attempt number	This is attempt 1.
Submission status	Submitted for grading
Grading status	Graded
Due date	Wednesday, October 28, 2020, 11:00 PM
Time remaining	Assignment was submitted 2 hours 22 mins early
Last modified	Wednesday, October 28, 2020, 8:37 PM
File submissions	Sam_Levin_HW_#2.zip October 28 2020, 8:37 PM
Submission comments	Comments (0)

Edit submission Remove submission

You can still make changes to your submission.

Feedback

Grade	94.00 / 100.00
Graded on	Friday, November 13, 2020, 9:13 AM
Graded by	 Shuo Niu