COGS9: Introduction to Data Science
Final Project
Due date: 2024 December 12 23:59:59 (Thursday)
Grading: 10% of overall course grade. 40 points total.
Completed as a group. One submission per group on Gradescope.

Group Member Information:

Please read the COGS 9 team policies to best understand how to approach group work and to understand what the expectations are of you in COGS 9.

First Name	Last Name	PID
Ryan	**Rulkens**	**A18619264**
Kanishk	**Hari**	**A18534775**
Ryan	**Cohen**	**A18627137**
Dylan	**Louie**	**A18615626**
Shaun	**Israni**	**A18361110**

Question (2 pts)

Clearly state the specific data science question you're interested in answering. This question can be the same as what you submitted for your project proposal. Alternatively, you can edit your original question or change your topic completely.

How can different road conditions predict the rate and severity of traffic accidents in the United Kingdom?

Hypothesis (2 pts)

Write down your group's hypothesis to your question. Provide justification how you came to this hypothesis. (What background information or instinct led you to that hypothesis?). You should incorporate the feedback you received on your proposal.

In general, crash rates and fatality rates will increase with poor road conditions, like snow, rain, or wetness. However, snow would lead to more crashes but less deadly results, as cars would be slowed down because of this direct interference, leading to milder collisions overall. On the other hand, frosty and wet road conditions would lead to more deadly accidents because of skidding and swerving from drivers oversteering. Additionally, when it's darker outside there

would be a lower quantity of crashes because less people are driving; however, there would be a higher proportion of crashes overall due to lower visibility.

Background Information (3 pts)

Include a few paragraphs of background research and information on your topic. This should include at least 2 citations to work from others. Including hyperlinks to reputable sources are fine.

For the dataset to test our hypothesis, we chose the UK specifically because its characteristics significantly lower the prevalent biases in accident rates when isolating road conditions and light levels, which makes our conclusions from the results more reliable because bad drivers cause a lot less accidents.

This is because the UK requires a higher level of driving education and skills since the driving test is much harder with a 40% pass rate in the UK compared to a 80% pass rate in the U.S. to get a license, so UK drivers are less likely to crash compared to other countries because of their higher driving standards. Additionally, the UK has a high percentage of single lane carriageways at 74%, which means that since there's only a single lane of traffic going in each direction there isn't much driver interference and thus less crashes are caused by others as well. In fact, this is represented through the fact that the UK is regarded as one of the safest countries to drive in. According to a study published by the UK Department of Transport, the UK ranked 5th out of 38 countries with available data for lowest number of road fatalities per million population. Despite data from Macrotrends reporting an increase of 8.1 million people from 1979 to 2022 for the UK population, the UK had a general downward trend in fatality rates since 1979, showing drivers are getting better and accidents are not due to reckless driving. Additionally, based on the demographic perspective cited for fatal alcohol crashes related to time of day, during midnight to 3 am is the prime time and the only one in which over 50% under alcohol influence actually crash for all drivers under the influence. But with Britain's road nature of a higher percentile of single lane roads, this involves crashing into another driver going the same way as one completely out of the way.

Although our study doesn't take into account other factors and variables like vehicle type (van/goods, bus/coach, motorcycles, and cargo trucks) and actual longitude/latitude of the crashes, which could all have certain impacts on the fatalities rates apart from just testing the weather's impact on road conditions and light amounts in areas, we can still observe that the difference with vs without those interferences is much less notable and significant than the main variable categories that we are testing in our project. The process on how we chose this dataset is based on how the dataset relates to our hypothesis and the background information aforementioned. And the variables also stated above that we would like to test the relationships between are clearly stated in the columns of our dataset, which is why we chose that one.

Lastly, we selected our dataset since it encapsulated the years from 2019 to 2022 - having the direct context of the pandemic influence which happened in those years in its entirety (except 2019, as it really started to affect other countries apart from China in January 2020). This pandemic variable could show us another dimension into how crash fatalities and

quantities changed as the pandemic was at its worst in near 2020 and got progressively better over time towards 2022 as well as how the crash scenario was like beforehand in 2019 when everything was normal.

Moreover, by choosing the UK we are preventing a lot of biases in general from the state and reputation of the country and the people, as shown in our dataset that we chose and will get more into. With our data, the government would technically have to assign serial numbers as reflected in the index to all the cars and track their crashes on highways and normal UK roads, so the source of the data is probably from them. Ultimately, bringing up the factors of road conditions, even though not weather conditions directly affecting like it being foggy or hazy, and light levels as reflected in our utilized dataset are really strong in establishing a direct comparison between environmental conditions, outside of the UK driver's or passengers' own fault (which wouldn't be much if anything because of the driver education and not much drinking problems), and this reinforces our approach and reasoning into choosing this question, dataset from kaggle, and trajectory in analyzing the data for our study.

https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2022/reported-road-casualties-great-britain-annual-report-2022

https://www.brake.org.uk/get-involved/take-action/mybrake/knowledge-centre/uk-road-safety

Time of Day and Demographic Perspective Of Fatal Alcohol …

Data (2 pts)

Include a description of the perfect dataset you would need to answer this question. How many observations would you need? What variables would you collect? Explain the perfect dataset that you would want to answer this question.

An ideal dataset would contain specific indication of brightness, ideally as a quantitative variable, as well as various indicators of weather factors, like wind, speed, rain, temperature etc. In addition, it would need a column describing the dampness of the road itself. We would need observations spanning multiple years to avoid biases from temporary conditions like COVID, as well as enough observations for categorical variables like rain vs snow vs sun to lower standard error of results. An ideal dataset would also have specific labels to allow for clearer interpretation of results.

Then, look online for available datasets. Find a dataset that could be used to answer this question. Describe how many observations are included and what variables have been collected. Discuss the dataset's limitations and how it differs from your ideal dataset. If you collected your own data, explain what information you collected, from whom you collected it, and a link to the data.

https://www.kaggle.com/datasets/nezukokamaado/road-accident-casualties-dataset

There are 14 variables in the dataset, which comprises numerical and categorical variables. The variable names includes: index (unique accident identifier), accident severity (ranging from slight, serious, and fatal), accident date, the latitude and longitude coordinates of the accident, the district where the accident occurred, the light conditions, number of casualties, number of vehicles involved, road surface conditions, road type, the area of the accident (urban, rural, unallocated), weather conditions, and vehicle type. There are around 660,680 rows in the dataset. Some limitations occur with the number of years the data was collected, as it was only collected between 2019 to 2022, so it does not account for crashes that have happened recently or beyond 2019. This differs from our ideal dataset as we would like our dataset to include the most recent crashes on a daily basis, which would allow us to analyze the results of our data.

Ethical Considerations (3 pts)

Read the data science ethics checklist from lecture. Then, discuss what ethical considerations must be made when answering your specific data science question. Brainstorm and explain how you would address these considerations for each of the following categories in your specific project: Team Bias, Sampling Bias, Data Bias, Consent, Data Privacy / Ownership, Algorithmic Bias / Discrimination, Transparency, Unintended Consequences, Continued Monitoring / Accountability. Feel free to write about additional ethical considerations you would make that aren't included on the checklist. Note that data privacy is NOT the only ethical consideration for a data science project. It is a piece, but there is a lot more that has to be considered.

For any data science project, it is critical to consider the ethics of the project's research at every stage. When generating our data science question, our group decided to choose to focus on the United Kingdom so we could avoid biases in many steps of research. Firstly, our group has little to no previous knowledge of how road and light conditions affect traffic rates across the world, and especially those in the United Kingdom. This protects against confirmation bias since no team member has a preexisting belief or hypotheses on the subject. Additionally, choosing the United Kingdom lowers sampling variability as the United Kingdom provides a large dataset of individuals who are relatively good drivers, which helps to avoid some bias in the data bias since the accidents are less likely to be caused by reckless driving.

Ethical consideration must also be made to avoid algorithmic bias and discrimination. Since the dataset avoids using any personal information about the drivers involved in the accident, such as sex, age, race, sexuality, and religion, algorithmic bias will arise from that. Additionally, the dataset includes accidents from areas across the United Kingdom so the data collection can avoid bias due to taking samples from only one area. To maintain this unbiased dataset, we must also ensure the data is continuously monitored and updated to its most recent form to address evolving data and ensure the dataset remains relevant to our hypothesis.

As previously mentioned, the dataset avoids using any personal information about the drivers involved in the accidents. Therefore, consent from each individual to use their data and the removal of data upon request does not need to be considered since the data is generic, but a system will be put in place to remove data upon request. If the dataset were to include personal information though, our team would need to obtain consent and be able to remove data upon request. However, we would need to monitor terms of use of the dataset we are using and ensure we are able to use the dataset through web scraping. Regardless of if the data contains personal information or not though, our team would have to constantly monitor who has access to the data and think about how the data could be used in a harmful way.

One ethical consideration that our dataset fails to address is bias in the data itself. While a lack of detail in regards to gender, age, race etc. improves the efficacy of our model, the dataset also fails to collect contextual details relating to the circumstance of crashes, thereby creating the possibility for wrong interpretation of results. For instance, a majority of crashes that occur between the hours of 12 and 3 am involve inebriation due to alcohol, thereby resulting in our dataset vastly overestimating the dangers of simply driving at night sober. We will have to uncover all forms of data bias in order to properly contextualize the results of our project and avoid unintended conclusions.

Analysis Proposal (15 pts)

Here, you will propose how you would use and analyze data to answer your question(s) of interest. You are neither expected nor encouraged to carry out the analyses to answer your question(s). You will describe, in detail, what you would need to do to prepare your dataset for analysis (data wrangling) and what type of analysis you would do to answer your question(s). Explain how your proposed methods / approaches would allow you to interpret the results from this analysis. We are looking for the correct conceptual understanding and application of ideas discussed in class, not specific and technical implementations. For example, if you are applying machine learning to some categorical data, it's important to specify whether you will be performing regression or classification. If you are unsure about the details of anything above, ask on Piazza, come to office hours, and/or do further research on your own (Stack Exchange, Google, Wikipedia, etc.).

Specifically, you are required to incorporate *at least four different methods*, exploring ideas from a combination of:

- Data Collection (web scraping, APIs, etc.)
- Data Wrangling
- Descriptive & Exploratory Data Analysis (summary stats, correlation, etc.)
- Data Visualization
- Statistical Analysis (Inference, A/B testing, etc.)
- Predictive Analysis (machine learning, classification, regression, etc.)
- Text Analysis (Sentiment Analysis, TF-IDF, etc.)
- Geospatial Analysis (choropleth maps, geospatial statistics, etc.)

Method 1: Data Collection

- Data Collection
    - The first part of analysis begins with collecting good data. It is necessary to find a proper dataset that can answer each part of our question. Data is able to describe both the different road conditions and severity of accidents. Finding a proper dataset ensures that we do not need to waste time and resources going back and finding more data. To avoid data that can be easily skewed or biased, we need to find a data set large enough (Our data set has over 661k entries). After going through our dataset, we have reviewed and confirmed that ours additionally does not bring up any ethical issues, values simplicity and effectiveness, has a strong data source, and had a successful data collection process/retrieval overall.

Method 2: Data Wrangling & Creating Tidy Data

- Data Wrangling
    - Currently, our data is formatted as a singular dataset containing numeric and categorical variables. However, the way the data was collected was a little untidy and unorganized; thus, by reorganizing the dates and columns/variables in which they put their data in to be more structured, specifically labeled, and in a favorable formatted order, we can tidily wrangle our data to achieve the most desired state of preparation for what we want to get out of our data. In addition, the year column is formatted incorrectly, currently being in the order DD-MM-YYYY; to fix this, we would use the Pandas library to format that column to YYYY-MM-DD.

Method 3:

- Data Visualization
    - We will create a bar graph about the number of traffic accidents per year in the UK and then we will create isolated bar graphs about the rating of how severe the road conditions are on average throughout the different years, and these ratings will be about snow and sleet, frosty and wet, and darker road conditions. This is enough to help answer our question by showing how road conditions can predict the rates of traffic accidents in the United Kingdom.

Method 4: Predictive Analysis

- Predictive Analysis

    We would need to develop two different models to predict the rate and severity of accidents in the United Kingdom. To predict the severity of accidents, we will use classification models. This works well for our hypothesis as the model will show how our input variables (such as light and road conditions) will affect the severity of traffic accidents. To train the model, the dependent (target) variable will be accident severity

(such as slight or serious). The independent (input/predictor) variables will be light conditions, road surface conditions, weather conditions, time of day, etc. In order to find a relationship between our factors and the severity, we will use a decision tree, wherein each independent variable will provide a weight to the resulting classification, outputting a vector of probabilities indicating the likelihood of each severity level occurring. To predict the rate of accidents, a regression model will be used. This works well for our hypothesis as the model will show how our input variables (such as light and road conditions) will affect the rate of traffic accidents occurring under different road conditions. To train the model, the dependent (target) variable will be accident severity (such as slight or serious). The independent (input/predictor) variables will be light conditions, road surface conditions, weather conditions, time of day, etc. Furthermore, we will ensure all independent variables have a numerical value, meaning categorical data like road surface conditions, weather conditions, and light conditions will be assigned a numerical value. In order to find a relationship between our factors and the accident rate, we will use linear regression, taking in the converted independent variables and outputting a single probability indicating whether an accident will occur or not.

Discussion (10 pts)

Given your hypothetical results, how would you draw inferences / conclusion based off of those results? What are the limitations, pitfalls, and potential confounds of your methods, or biases in your data sources (e.g., how does the selection of the sources of your crowds affect your outcomes?)? How would you set out to address them? In addition, outline how you would address any societal and/or ethical implications of your proposed project discussed in your Ethical Considerations section.

The conclusions we would draw from our results would be the key factors that affect the severity of accidents in the United Kingdom, and the probabilities of accidents occurring under various road conditions. To do this, we would utilize various different algorithms, such as decision trees, to show how much conditions like snow contribute to severe accidents, and what measures can be taken in order to mitigate its effect on people. Some limitations in our data source include a lack of data in rural areas or certain weather conditions (example: hail), which can affect generalizability if the data is biased towards conditions such as rain and sunny. A potential confound we may have is that, when applying algorithms, there may be some overfitting, which can cause our data to show unrelated relationships and not capture complex interactions. To address these issues, we will break up our data into test data and training data to ensure validity. We would also examine interaction effects between variables to potentially capture non-linear relationships between these variables. Some ethical considerations in our project would be to make data confidential and private, so that parties who were a part of the accidents in the data will not have personal information publicly available. The underlying message in the project would be to highlight actions that the United Kingdom government can take to mitigate the rate of accidents occuring, such as driving in unfavorable weather. Finally,

we would communicate our findings in a way that dismantles any chances of biases or oversimplification being displayed.

Group Participation (3 pts)

Include one paragraph briefly outlining the contribution of each group member throughout the quarter while working on this project. Each of you must also fill out the survey (link provided toward the end of the quarter) about individual and group participation. **The results of this survey can negatively impact an individual's final grade if the group provides evidence that one member did not contribute to the project**. (3 pts)

> During the progression of this project, each group member participated on an equivalent level. Each member was involved with the development of our question, hypothesis, finding the dataset, and providing feedback to other member's sections. Each individual though, did complete a portion of the project by themselves. Shaun Israni was involved in the prior research for the question, constructed the hypothesis, and worked on the background information with Dylan as well as helped with the data visualization and wrangling methods. Dylan Louie worked on the background information with Shaun and also data collection under the analysis section. Ryan Rulkens worked on the data, ethical consideration and hypothesis for the initial proposal, as well as the data wrangling and predictive analysis sections in the final project. Ryan Cohen worked on ethical consideration for the initial proposal, as well as predictive analysis sections in the final project. Finally, Kanishk Hari worked on the data and discussion sections of the project, ensuring that the data for this project was sufficient to meet all the criteria required for the project to succeed, and that discussion was thoroughly completed to answer unanswered questions.