COMP5310 PROJECT STAGE 1

Explore, clean, summarise and analyse the data

Group Name: Group 9
Activity Number: 15
Group Number: 9

Student ID	UniKey	Report ID from Table of Contents
540348673	aaln0584	I
540371073	nisa0298	II
540349681	kpoo0923	III

MARCH 27, 2024

THE UNIVERSITY OF SYDNEY
Camperdown NSW 2050

Table of Contents

SID: 540348673**, UniKey:** aaln0584

I. Airbags May Cause Injuries

A. Research Question

Airbags are designed as a compulsory system in all cars; however, this was not the case until early 90s [1]. They provide safety for passengers in case of a car crash as they activate to support the driver or the passenger. While they do provide safe measurements, they might also do harm as well. In this report, we investigate the question “Does airbag deployment increase the injury severity in a car crash?”.

B. Shareholders

The shareholders are automobile manufacturers, insurance companies, regulators, emergency medical units, and consumers. Manufacturers may revise why injury severity increases with airbag deployment and make changes that mitigate it. Insurance companies can factor in the injury cost in car crashes differently when an airbag deploys. Regulators may set stricter guidelines for airbags system to help in lowering injury risk. Emergency medical employees could have more caring protocol for crashes that have a deployed airbag. Finally, consumer may decide to buy a more trustworthy car based on the reliability of airbag system and brand.

C. Data Description

The National Highway Traffic Safety Administration (NHTSA), Agency of USA federal government, keeps track of traffic crashes injuries. The data is collected by the police who reported the incidents, and it covers 1997-2002 car crashes events, and it only includes a subset of the variables and observations. The data has 16 attributes and 26217 observations. The full details of the data are in [1Data structure and metadata].

1. Provenance of the data

The data is reconstructed and combined by Dr. Mary C. Meyer, Tremika Finney, and corrected by Dr. Charles [2]. According to them, the data is publicly available on the NHTSA site, and they merged multiple datasets to create it. I obtained the data from vincentarelbundock Rdatasets [3], which has more than other 2000 datasets from R.

2. Data license

The data originally is available under public domain license without the need for citation [4][5][6]. However, the reconstructed data belongs to Dr. Mary, Tremika, and Dr. Charles [2], and it should be sourced accordingly since they did not mention the exact license.

D. Data Quality & Cleaning

I did the data processing in python, using Pandas and other libraries. The data is constructed in csv format, hence tabular data that can be accessed by row and column. The values in the data are mostly categorical with different values, around 2-3 values for each category. Furthermore, it has some float, integer, and date variables.

1. Unique values

I checked the unique values of each column using Pandas unique function to find any inconsistent values. Impact_Speed column has inconsistent levels, so I fixed them, and I created a new variable that uses the range of Impact_Speed to generate normally distributed random data of cars speed. I have chosen normal distribution because it is more suitable for real-world natural phenomena than, say, random integer. It will be useful later for analysis and data exploration.

2. Treat nulls

I filtered the null columns using Pandas isnull() function and I found two columns that have null values. yearVeh column had 1 missing value. I decided to fill it with the average value because both the range of years and standard division are small, hence the average value should be suitable. As for the second column, it was InjSeverity, and I decided to keep them because they might be useful for exploration. They can be filtered out easily when needed.

3. Drop irrelevant columns

I dropped a few columns that has no use for me using Pandas drop function. First one is “rownames”, and it was basically just an index column. Second, I dropped “caseid” because it has no use for me as it was just a car identifier. Third, I dropped “abcat” because I found that it is just created using the combination of airbag and deploy columns.

4. Update data types and rename columns

I changed few columns data types to be suitable for data exploration using Pandas astype function. Furthermore, I renamed the columns names for convenience using pandas rename function as indicated in [Data structure and metadata].

5. Make values relevant

I decided next to work on the inconsistent data values. I used the code in Figure 1 to change them to “Yes” and “No” responses. This helps when doing multiple graphs since the names are mostly the same for all variables.

```
df['Deceased'] = df['Deceased'].apply(lambda x: "Yes" if x == 'dead' else "No")
df['Avail_Airbag'] = df['Avail_Airbag'].apply(lambda x: "Yes" if x == 'airbag' else "No")
df['Seat_Belt'] = df['Seat_Belt'].apply(lambda x: "Yes" if x == 'belted' else "No")
df['Front_Impact'] = df['Front_Impact'].apply(lambda x: "Yes" if x == '1' else "No")
df['Role'] = df['Role'].apply(lambda x: "Driver" if x == 'driver' else "Passenger")
df['Airbag_Deploy'] = df['Airbag_Deploy'].apply(lambda x: "Yes" if x == '1' else "No")
df['Sex'] = df['Sex'].apply(lambda x: "F" if x == 'f' else "M")
```

Figure 1 - Code to change the values

6. Filter Veh_Year variable

The first car that had an airbag compulsory to its system is the Porsche 944 in 1987 [1]. Therefore, I decided to filter out the observations that have cars manufactured before this period. I removed 3424 observations, which left 22793 rows. The data exploration analysis should give a better result now, even though we have lost some information.

E. Exploratory Data Analysis (EDA)

1. Injury Severity exploration

First, I decided to explore Injury Severity column and compare it to the others. It has a range of 0 to 6, standard deviation of 1.3, median of 2, average of 1.7, and most of the data are concentrated on the first half. I made my decision to report the average because the median is only going to report 6 integers, mostly between 1-3, which are not suitable for comparison. I compared Avail_Airbag, Seat_Belt, Front_Impact, and Airbag_Deploy, all vs Injury Severity. The result is shown in Figure 2. From the result, we can see that airbag deployment increases injury factor, seat belts lower the injury factor, and having an airbag in the car is better than none. The increased mean value of injury severity due to an airbag deployment is shown in Figure 3.

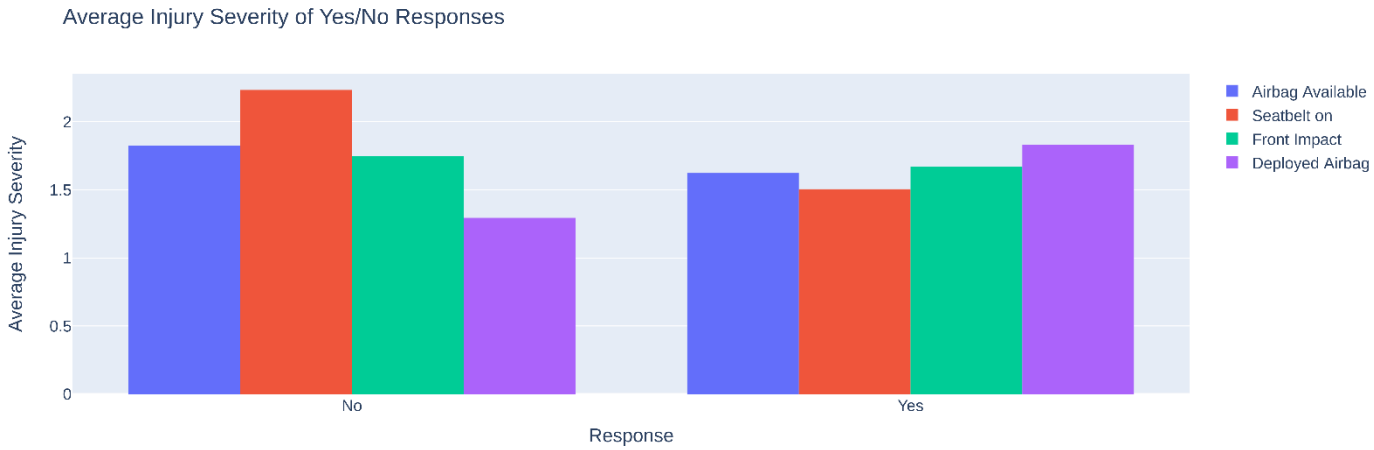

Figure 2- Average Injury Severity Response

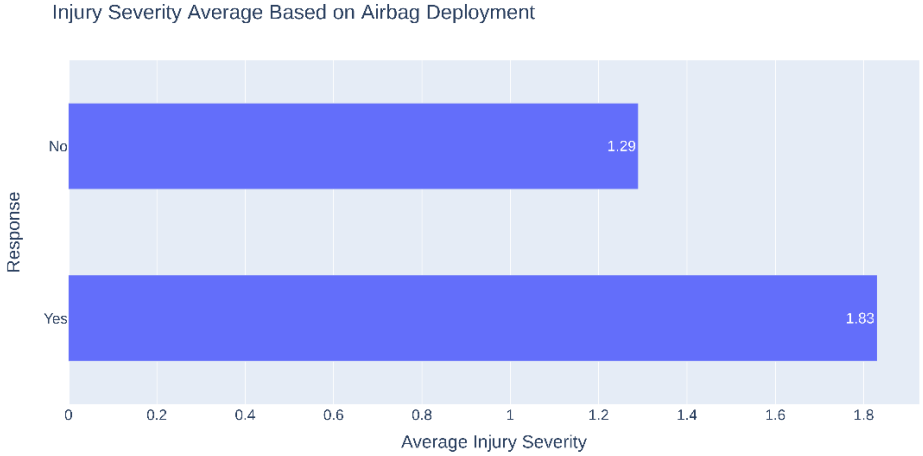

Figure 3 - Injury Severity vs Airbag Deployment

Furthermore, I decided to investigate the numerical variable Impact Speed in Figure 4. Impact speed seems to be a determinant factor of fatalities and injury severity as it has higher value when people are dead from a car crash.

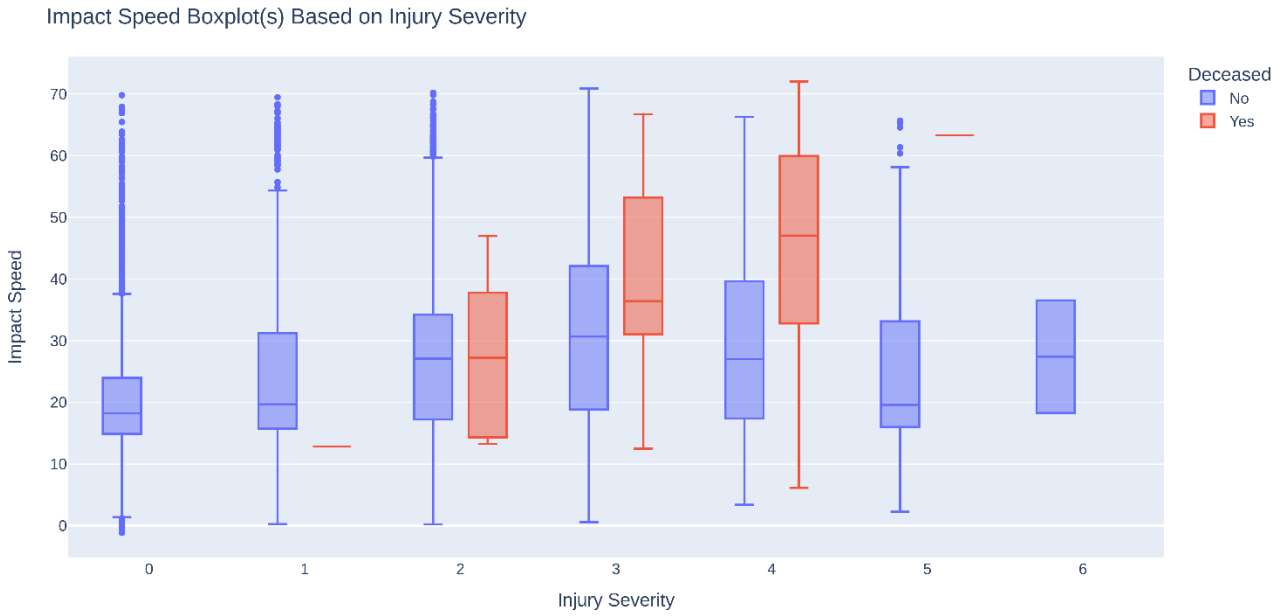

Figure 4 - Impact Speed vs. Injury Severity Boxplot(s)

2. Data distribution

I found that 37% of cars in the data do not have airbags available to deploy. Even after cleaning the data by vehicle year, there is still many cars with no airbags. I kept that in mind, and I made my analysis with cars that have airbags. Next, 26% of drivers or passengers did not wear seatbelt. These might be the ones that died or suffered the most during the crash. Moreover, 61% of the cars that have an airbag during a crash had it deployed. It means that around 39% of the crashes were not severe.

3. Investigate Weighted analysis

The data has included a useful column, which is the weight column. Weight column purpose is to account for sampling bias according to the data source [7]. Figure 5 shows that injury severity increases during an airbag deployment, which indicates that they might be correlated.

```
df_dep = deploy_df[deploy_df['Airbag_Deploy'] == 'Yes'].dropna()
df_nodep = deploy_df[deploy_df['Airbag_Deploy'] == 'No'].dropna()
weighted_stat = sm.stats.DescrStatsW(df_dep['Injury_Severity'], weights=df_dep['Weight'])
mean = weighted_stat.mean
std = weighted_stat.std
q = weighted_stat.quantile([0, 0.25, 0.5, 0.75, 1])
print(f"Injury Severity when an airbag has been deployed \n mean: {mean}, std: {std}, quantiles: {q}")

weighted_stat = sm.stats.DescrStatsW(df_nodep['Injury_Severity'], weights=df_nodep['Weight'])
mean = weighted_stat.mean
std = weighted_stat.std
q = weighted_stat.quantile([0, 0.25, 0.5, 0.75, 1])
print(f"\nInjury Severity when no airbag has been deployed \n mean: {mean}, std: {std}, quantiles: {q}")
```

```
Injury Severity when an airbag has been deployed
 mean: 1.1156150161860605, std: 1.1152067839375217, quantiles: p
0.00    0
0.25    0
0.50    1
0.75    2
1.00    6
dtype: int64

Injury Severity when no airbag has been deployed
 mean: 0.5328387488968972, std: 0.8740268501553905, quantiles: p
0.00    0
0.25    0
0.50    0
0.75    1
1.00    5
dtype: int64
```

Figure 5 – Airbag Deployment Weighted Analysis

F. Conclusion

The exploratory data analysis shows that seat belt, airbag deployment, and impact speed might have some correlation with injury severity or fatalities during a car crash. Furthermore, airbag deployment has been shown to be related to the increase in injury severity in multiple scenarios. The next step is to make a data model to confirm our claims, and see if airbag deployment increases the injury severity, and how the variables of interest affect the injury severity results.

II. Predicting NBA 3-Point Shot Outcome from Player-Agnostic Variables

A. Research Question

There are many players in the National Basketball Association (NBA) that are recognized as great 3-point shooters for their ability to maintain a high percentage of made 3-point shots throughout the season. These players are often highly sought after by general managers and become an expensive asset to have on a team's roster. However, it is difficult to distinguish how much of a shooter's success is due to their own skill versus the play-making and offensive strategy of the broader team. In other words, if a player achieves a high 3-point shooting percentage and is wide open on every shot, then it is possible that the same outcome could be achieved by an average player. Given a general manager's objective to minimize the team's cost base while having a winning season, understanding the true return-on-investment of a 'great' 3-point shooter is key. Thus, our research question:

Can 3-point shot outcome (make/miss) be accurately predicted from shooter-agnostic variables?

Understanding the shooter-agnostic variables (defender distance, period, dribbles[1] before shot, etc.) that may be predictive of shot outcome provides general managers, coaches, and team owners with the necessary tools to execute on a highly efficient roster. If we found, for instance, that the highest probability shots are those with medium defender distance and no dribbles, then the offensive motion and plays should be optimized to maximize those situations. Alternatively, if we found that shots made by certain teams are particularly sensitive to period in the game, then it is possible those teams have relatively weak stamina. Ultimately, having a clear understanding of these variables' influence allows for much more informed decision making. These are arguably the most economically efficient levers that can be pulled to achieve a winning season, as they are based on a fixed player roster.

B. Dataset

The dataset obtained consists of 128,069 shots from the 2014-2015 NBA season, along with 21 corresponding attributes. This dataset was downloaded from Kaggle on March 11, 2024, with no license restrictions and was originally sourced from the NBA's publicly available REST APIs. Please refer to Appendix B-1 for complete metadata and data dictionary. Intuitively, the most relevant attributes to our research question are those that may contribute to shot difficulty. This includes variables such as distance from closest defender, distance from the basket, dribbles, and time of possession. Other attributes, such as final margin, are more concerned with the broader game than the individual shot and thus irrelevant to our research question.

C. Data Ingestion & Cleaning

The dataset was processed as a Pandas data frame with each record corresponding to an individual shot. Given our focus of 3-point shots, all 2-point shots were removed, thus shrinking our data frame by about 75%. Data cleaning primarily involved type conversion, validation of data consistency, removal of outliers, and parsing of multiple attributes from a single column. As presented in Figure 1, some nominal attributes (e.g., PTS_TYPE, player_id) needed to be converted to strings. On the other hand, GAME_ID contained multiple pieces of information that needed to be split into individual columns. Leveraging *datetime.strptime()* and *apply()*, the stadium and game date were parsed out from GAME_ID. Stadium labels were validated by comparing the *unique()* output to our list of 30 teams. Additionally, the stadium variable allowed us to validate the distribution of home game counts across all teams. This also prompted reviewing the *min()* and *max()* game dates, as home game counts appeared below what was expected. Ultimately, we found that stadium labels were correct but that our dataset excludes the last 1.5 months of the regular season. This does not raise concerns in the context of our research question, as our sample size remains substantial. We will proceed under the assumption that shots in the last 1.5 months do not structurally differ from the rest of the regular season.

The player_id, player_name, CLOSEST_DEFENDER, and CLOSEST_DEFENDER_PLAYER_ID attributes were validated against each other, respectively, as follows:

```
(df.groupby('CLOSEST_DEFENDER')['CLOSEST_DEFENDER_PLAYER_ID'].agg(lambda x: x.nunique()) > 1).sum()
```

One player was found appearing under two different player_id's and was subsequently corrected. DRIBBLES was also converted from integer to boolean, as the total number of dribbles prior to the shot is irrelevant to our research question.

The final cleaning step involved identification and removal of outliers. Matplotlib boxplots were leveraged to understand the distribution of numerical attributes and outliers were initially detected based on 1.5 x IQR. Outliers for CLOSE_DEF_DIST and TOUCH_TIME were adjusted to 3.0 x IQR as the distributions appeared to have particularly long tails. Following removal of redundant or irrelevant columns, our clean data frame consists of 29,806 shots (records) and 12 attributes (including shot result).

D. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) primarily consisted of comparing the attributes of made versus missed shots as well as better understanding the relationships between the attributes themselves. More specifically, this stage was guided by the following descriptive questions:

1. Are there clear structural differences in shot outcome throughout the course of the season?
2. How much dispersion is there across shot distance and defender distance? How do the distributions of made versus missed shots differ along these two axes?
3. Are there any patterns between controllable offensive variables (e.g., touch time, dribbles) and defender distance?

It is possible that the relationships between our attributes and class (shot outcome) evolve throughout the course of the season. If our dataset included playoffs, for instance, we might see noticeable changes in our data resulting from increased pressure. Since we are strictly concerned with the regular season, however, there are fewer intuitive reasons for why we might see relevant structural changes in our data throughout the course of the season. Figure 1 below helps us answer this question by showing how average 3-point shooting percentage (across all players) fluctuates throughout the season.

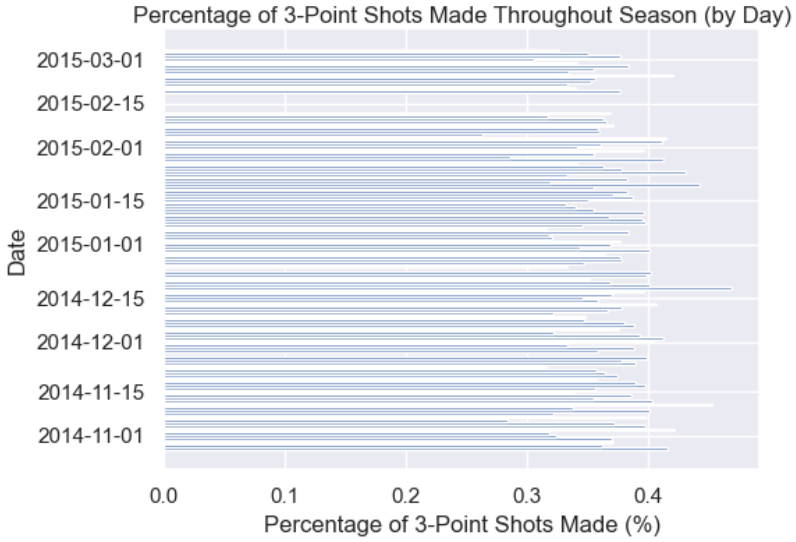

Figure 1: Mean Percentage of 3-Point Shots Made Throughout Season (by Day)

The bar plot shows a relatively stable 3-point shooting percentage with no clear patterns, which is consistent with its low (3.4%) standard deviation. Though it may require further validation during the modelling phase, we will proceed EDA under the assumption that there are no structural changes in our dataset over the course of the season.

We begin by focusing on shot distance and defender distance as these are (intuitively) the variables most directly linked to shot difficulty. As shown in **Figure 2**, shot distance appears approximately normally distributed and there is no noticeable difference between the distributions of made versus missed shots. Moreover, the standard deviation of 1.22 feet suggests that there is very limited dispersion in shot distance. In other words, even if shot distance was predictive of shot outcome, it would be very difficult to implement a game strategy that optimized for 1–2-foot changes in shot distance. Closest defender distance, on the other hand, shows both a moderate difference in distribution between made and missed shots (mean of 6.2 versus 5.9 feet, respectively) as well as greater standard deviation (~2.5 feet).

[1] Dribbling refers to when players bounce the ball; this is commonly done prior to taking a 3-point shot.

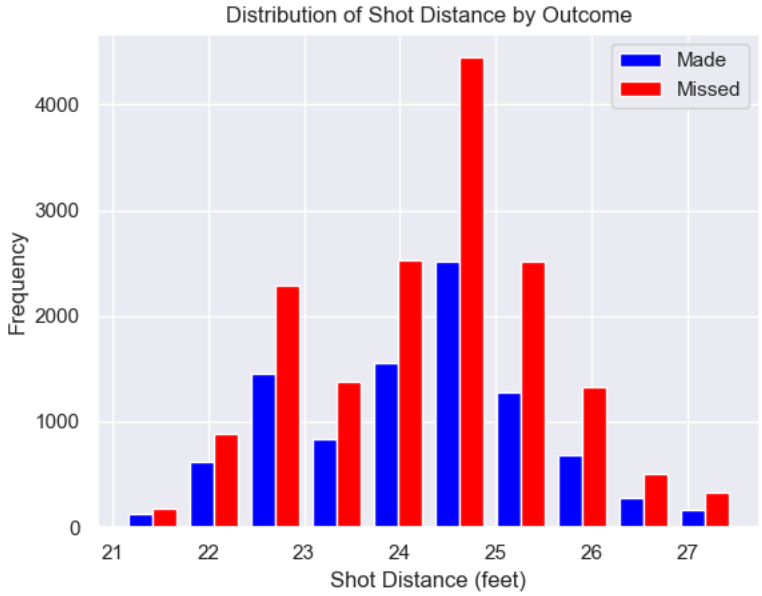

Figure 2: Distribution of Shot Distance by Outcome

The next step was to identify the typical characteristics of shots with greatest distance between the defender and shooter. As shown in **Figure 3**, for a fixed shot distance, there is some noticeable difference between defender distance for shots with dribbles versus shots with no dribbles. Specifically, as players get farther from the basket, dribbling tends to attract defenders more than not dribbling. If our objective is to maximize distance from the nearest defender, then one strategy might be to encourage less dribbling prior to the shot.

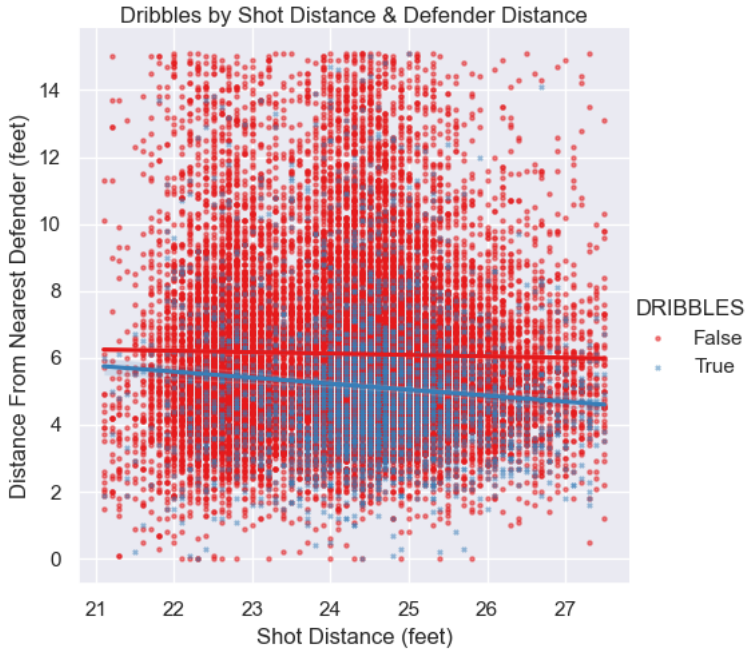

Figure 3: Dribbles by Shot Distance & Defender Distance

Another variable that may be at play in this interpretation is TOUCH_TIME. The average amount of time that a player holds the ball prior to shooting is 1.96 and 0.94 seconds for dribbles and no dribbles, respectively. This difference in possession time could allow defenders to approach and thus be closer when the shot is taken. A critical next step in the modelling phase would be to get a better understanding of the interactions between defender distance, dribbles, and touch time, as they appear closely linked. A clustering analysis may also help validate whether STADIUM or PERIOD are relevant to our research question, as they did not reveal any clear relationships in EDA. The GAME_ID was largely ignored during EDA, as there isn't an intuitive explanation for seeing how variables would change between games; however, this too may warrant further exploration in the next phase.

III. Likelihood Prediction for Term Deposit

A. Research Question

Since the global economy is currently experiencing a limp, world economic growth is projected to decline from 3.5 percent in 2022 to 3 percent this year and 2.9 percent next year. People are becoming more cautious in choosing their investment plans. Therefore, subscribing to a term deposit is receiving more attention due to its low risk but promising returns. In this report, we will investigate the question, ***"Predicting Likelihood of Term Deposit Subscriptions"***.

A term deposit is a type of investment where you put your money into an account at a bank or other financial institution for a fixed period. During this time, you agree not to withdraw the money. In return, you receive a higher interest rate compared to regular savings accounts. Term deposits help banks invest in other financial products that offer higher returns.

B. Data Description

Provenance of the data: The data was obtained from a marketing campaign conducted by a bank in Portugal. They gathered information by phone call and collected personal details from the people they called. The bank contacted each person multiple times to see if they wanted to subscribe to a term deposit. The institution then grouped the data collected into a csv file. "Bank-full.csv" dataset was downloaded on 19/3/2024 from https://archive.ics.uci.edu/dataset/222/bank+marketing, and it would be analysed in this report.

Data license: This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

Data Structure and metadata: The data originally had 17 attributes and 45211 instances. The metadata and data dictionary introduced the dataset by listing all information. [(Report 3) Appendix]

C. Data Quality and Cleaning

The analysis was conducted using Python in a Jupyter notebook. The Pandas and other libraries were selected to process the data and were imported to perform data cleaning and visualization. The file downloaded in csv format, which used to store tabular data. Column name: The variable name "y" was changed to "term_deposit" to provide a clear description (Code 1). This substitution aims to make clarification audience who may interact with it in the future.

```
[3]: df.rename(columns={'y': 'term_deposit'}, inplace=True)
```

Code 1: Use python code to rename the selected column

Dealing with missing value: The dataset had no missing values. However, when inspecting the dataset, all missing values were replaced with "unknown" data. Since most of these "unknown" values were from categorical data, it was not suitable to replace them with other data or use other approach, for instance, replacing the "unknown" with mean. To maintain data integrity and improve the prediction model, it was decided to remove rows containing "unknown" values (Code 2). Moreover, "contact" and "poutcome" had recorded over a ten thousand "unknown", to hold the completeness of the dataset, both columns would be removed.

```
[8]: df_new = df_new[~df_new.apply(lambda row: row.astype(str).str.contains('unknown').any(), axis=1)]
```

Code 2: Use python code to remove "unknown" value in each code.

Shrinking the dataset: Data reduction was implemented as the focus of this report is to predict an individual action. Hence, variables including "contact," "pday," and "poutcome" were removed from the analysis due to their lack of relevance to the study objective.

Changing the data type: To create a reliable predictive model later, it is important to convert categorical variables into numeric values. In the features "housing", "loan", "previous", and "term_deposit" where "yes" and "no" were represented, they needed to be replaced by “1” and “0”, respectively (Code 3). Converting them to numbers helps in performing mathematical calculations efficiently and simplifies the modelling process, resulting in a more convincing predictive model.

```
[9]: binary_mapping = {'yes': 1, 'no': 0}

     for column in df_new.columns:
         if set(df_new[column]).issubset(['yes', 'no']):
             df_new[column] = df_new[column].map(binary_mapping)


     df_new.head()
```

Code 3: Use python code to replace “yes” and “no” with 1 and 0

Extracting date variable: To complete the data cleaning process, "month" and "day" variables were used to combine into a single "date" variable (Code 4). Since both variables were originally in different formats and did not hold significant research value on their own, combining them into a date variable would be necessary.

```
[11]: df_new['date'] = df_new['month'] + ' ' + df_new['day'].astype(str)

      df_new.drop(['month', 'day'], axis=1, inplace=True)
```

Code 4: Use python code to combine “day” and “month”

After cleaning and modifying the dataset, the pre-processed dataset contains 13 columns and 43193 rows with no “unknown” value and easy interpretation of variables’ name. The dataset is now ready for exploratory analysis.

D. Exploratory Data Analysis (EDA)

In this report, Exploratory Data Analysis (EDA) was used to identify distinctive features of individuals within the dataset. The aimed to predict the likelihood of client’s behaviour; thus, EDA focused on discovering relationships between different attributes by examining graph patterns. By analysing individuals’ background like loan history, bank balance, job type, etc, patterns could be explored, and assumption could be made. Through visual exploration, the goal is to find key variables that appropriate for implementing to a likelihood model.

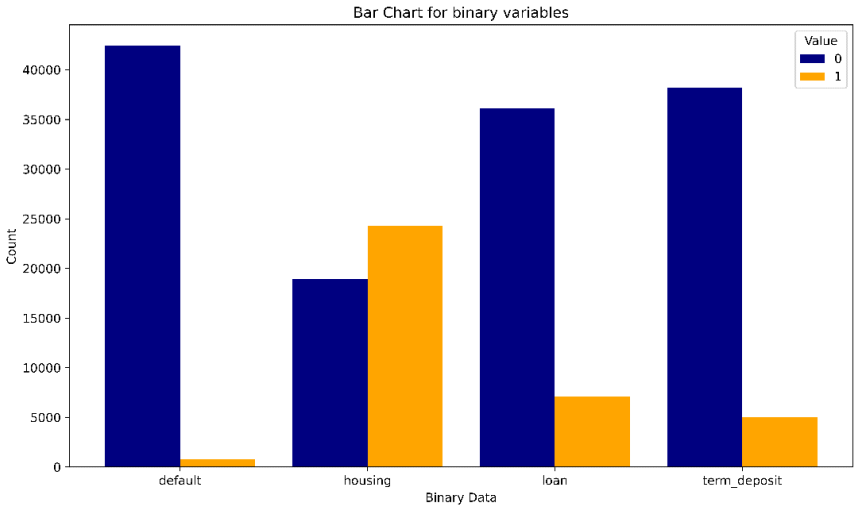

Figure 1: Bar chart for all binary variables

Figure 1 indicated the possible liabilities that an individual might have. It suggested that most participants do not have default credit history or personal loans, yet over half of them report having housing loans. Notably, a substantial portion of participants opted not to subscribe to a term deposit. Based on this graph, it seemed that personal liabilities may not be significant predictors for term deposit as we could not observe any decisive result.

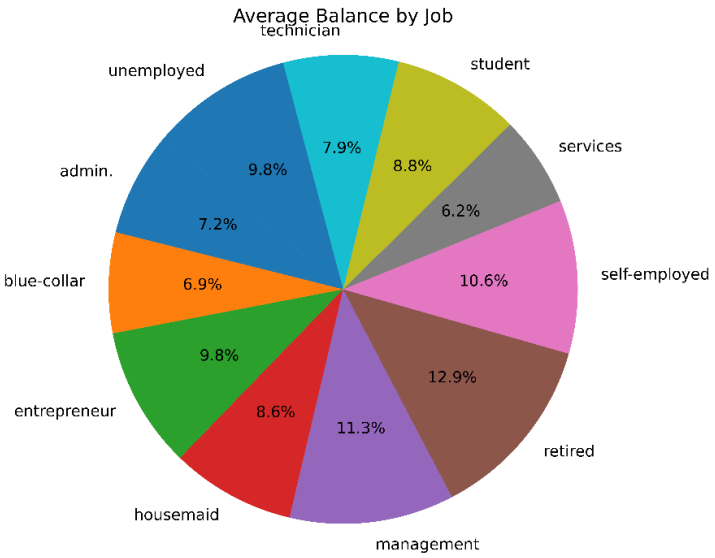

Figure 2: Pie chart for distribution of average balance by job

Figure 2 gave us an insight into the banks' balance an individual from different jobs typically had. It showed that retired people had the most balance on average, while managers and self-employed individuals had over 10% each. On the other side, those in service jobs and blue-collar workers had the least balance. Surprisingly, students had more money on average than admins, technicians, service workers, house maids, and blue-collar workers. Based on these findings, we might make an assumption that jobs like management, retirement, and self-employment tend to have a higher balance, making people more likely to spend for a term deposit. Conversely, jobs in services, blue-collar work, and admin roles might not pay as much, making them less likely to opt for a term deposit. It showed that average balance and job can be significant factors for predicting the likelihood.

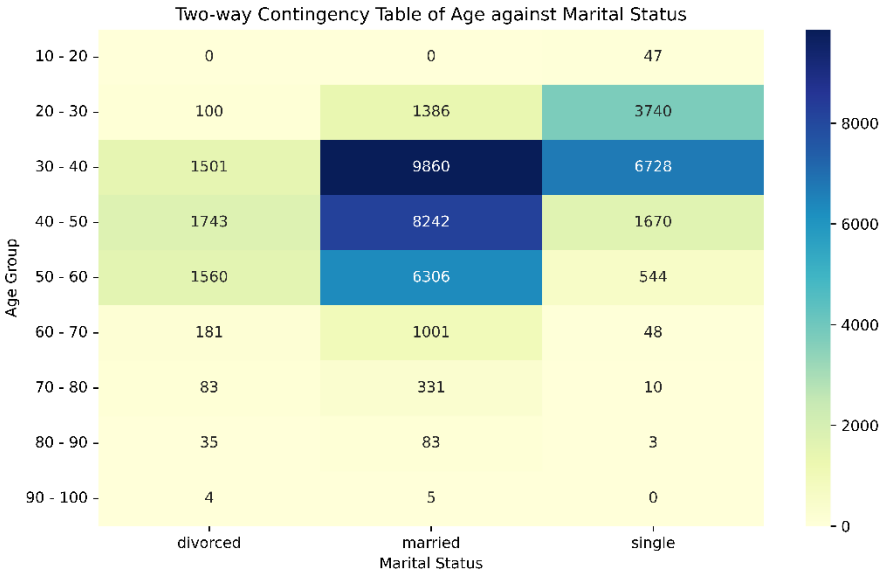

Figure 3: Two-way Contingency Table with job against education

Figure 3 showed a table with the age group and marital status of the participants. “Age group” was created by dividing “age” into to nine groups with a fixed range. It revealed that the majority of divorced and married people were in the age groups of "30 - 40," "40 - 50," and "50 - 60." Single individuals were mostly in the "20 - 30" and "30 - 40" age groups. Overall, it seemed like most of the people in the study were married and in the working age range. It indicated that both variables have uneven class, that “married” and “30 - 40” are the vast majority of variable “marital” and “age”, including these variables in the likelihood model might cause bias model, therefore, should be considered remove in the next stage.

IV. Group Component

A. Discussion

1. Airbags May Cause Injuries

As for strengths, the dataset has many variables that relate to the research question. Impact speed, airbag deployment, airbag availability, and injury severity, as discovered through EDA, appear to be most relevant. Furthermore, the dataset includes over 25,000 observations, which adds confidence to the conclusions reached during EDA. Nonetheless, there are apparent limitations. First, the data is from 1997 and may be outdated in the context of our research question and business need. Furthermore, the data lacks geographical attributes, driver or passenger medical history, car model and brand, airbags sensor failure rate, crash severity, and impact direction. Geographical data may help to identify if there are particularly high-risk locations for car crashes that should be investigated, while car model and airbag sensor failure rate may help in investigating the cars that have system malfunctions and show abnormal results in our analysis. In the EDA section, a visible strength is that it discusses the impact of airbag deployment vs injury severity, and it shows that it increases through different analytical methods during a car crash by comparing the mean. However, it is likely that the pattern seen between airbag deployment and injury severity has a confounding variable: crash severity. For us to reach meaningful conclusions on the causality of airbag deployment, we would need to isolate the effect by including a proxy for crash severity. This proxy is partially obtained through the Impact_level attribute, though it may not be exhaustive. The EDA is limited in that aspect as it does not study the Impact_level attribute when it is in critical value, level 3 or more.

2. Predicting NBA 3-Point Shot Outcome

The approach taken to predict 3-point shot outcomes from player agnostic variables is aggressive, as there are likely many factors, from fans and player health to teammates and coaching dynamics, that influence our class. As-is, the dataset is primarily limited in terms of attributes. Though the original dataset includes 21 columns, there is considerable overlap in the conveyed information. EDA suggested that any predictive model would likely be heavily reliant on CLOSE_DEF_DIST, as the other attributes, on their own, did not show significant changes between made and missed 3-point shots. Nonetheless, one strength of the dataset is that it is extensive and presents an opportunity to create new attributes from aggregating records. Specifically, one alternative approach is to bring in attributes that help segment shots by player type. Shot attempts from shorter players, for instance, are likely more sensitive to defender distance. Such attributes could be incorporated in various ways. The most obvious option is bringing in 3rd party data that we can link to players in our existing dataset. Another option would be to leverage the ~100k records of 2-point shot data that we originally excluded. With this data, we have an exhaustive picture of shots taken during the season. We could then use a clustering model to group players by shot types (this would potentially be more accurate than simply adding a position attribute). This grouping could then be leveraged in our predictive (shot outcome) model. We may end up steering away from the *player agnostic* component of our research question, however, if the groupings overfit our player data. This could be prevented by limiting the number of groupings during clustering.

3. Predicting Likelihood of Term Loan Deposit Subscriptions

The dataset gives a well-rounded description of the background on the individual, including details such as job type, age, marital status, and loans. One strength of the dataset is that it includes a diverse set of attributes with generally high variance. This allows us to explore many hypotheses and gives us the potential to explain a significant amount of variance in our class. However, there are also quite a few limitations, ranging from data quality to predictability. The "date" variable lacks information about the year, making it less useful and inappropriate for inclusion in our analysis. Most importantly, there are questions around the actual relevance of our attributes. That is, the attributes are largely demographic and do not include critical details, such as income and employment history. Based on the plots and table from our EDA, we found no indication that an individual's liabilities process any predictive value for term deposit. Most importantly, the average bank balance and the type of job are reliable indicators for predicting the likelihood of subscribing to a term deposit. The average balance indirectly reflects a person's financial situation, which can impact their decision to subscribe to a term deposit. It is also worth noting that our dataset is limited to a minority subset of subscribers who enrolled through the campaign, so our understanding of subscribers is biased. To obtain more accurate results, we suggest focusing to the average yearly balance and exploring its relationship with other background attributes.

B. Conclusion

To conclude our evaluation, we recommend moving forward with Report I (Airbags May Cause Injuries) as the dataset includes a good balance of attributes and EDA showed potential predictability. The feature quality is also generally high, as compared to the Report III (Term Deposits) dataset. Furthermore, the factors are measurable, which translate for quantifiable modelling and more interpretable results compared to Report II (NBA 3-Point Shot). Results for Report III showed that people who are in management, retired, or self-employed may have a moderately higher chance to take a term deposit. Report II (NBA 3-Point Shots), on the other hand, presented a strong business need and relatively high-quality data. During EDA, we investigated different approaches to manipulating defender distance, as this appeared to be the variable most related to shot outcome. The results indicated a relatively weak relationship between dribbles, shot distance, and defender distance. Thus, a predictive model for this topic would likely require additional attributes describing both players and match circumstances. Finally, Report I (Airbags May Cause Injuries) showed promising indications of a relationship between airbag deployment and injury severity, primarily by comparing the distributions of injury severity in both cases.

Appendix

A. Airbags May Cause Injuries (Report 1) Appendix

1. Data structure and metadata

Metadata	Information
Title	Airbag and other influences on accident fatalities
Description	US cars crash data (1997-2002)
File format	csv
File size	2.29mb
Details/Keywords	police-reported, harmful events, front-seat
Methodology	Multistage Sampling
Contact Information	meyer@stat.colostate.edu
Organization	Colorado State University
Language	English
Number of attributes	16
Number of instances	26217

Variable	Updated Variable Name	Description	Value after Processing	Type before processing	Type after processing
dvcat	Impact_level	Categorical with estimated impact speed level. It takes 5 fixed values, where each indicates the range of impact speed in km/h	Same as original	Object	Object
weight	Weight	Observation weights to account for sampling probabilities	Max-Min Positive Normalization	Float64	Float64
dead	Deceased	Categorical with dead or alive indicator	Yes/No	Object	Object
airbag	Avail_Airbag	Categorical with none or airbag if a car has an airbag	Yes/No	Object	Object
seatbelt	Seat_Belt	Categorical with none or belted if they have a seatbelt locked-in during the crash	Yes/No	Object	Object
frontal	Front_Impact	Categorical with 0 or 1 for front impact	Yes/No	Int64	Object
sex	Sex	Categorical with f or m for sex type	F/M	Object	Object
ageOFocc	Age	Age of occupant	Same as original	Int64	Int64
yearacc	Acc_Year	Year of accident	Date	Int64	Datetime64
yearVeh	Veh_Year	Year of vehicle	Date	Float64	Datetime64
abcat	Deleted	Categorical variable that answers if an airbag has been deployed. It has the values deploy, nodeploy, and unavail.	-	Object	Deleted
occRole	Role	Occupant role. It has the values driver and pass.	Driver/Passenger	Object	Object
deploy	Airbag_Deploy	Categorical with 0 or 1 for airbag deployment.	Yes/No	Int64	Object
injSeverity	Injury_Severity	A multi-level numeric vector for injury severity. 1: possible injury, 2: no incapacity, 3: incapacity, 4: killed, 5: unknown, 6: prior death.	Same as original	Float64	Int64
caseid	Deleted	Unique identifier for the vehicle.	-	Object	Deleted
New_Variable	Impact_Speed	Random normal distribution based on Impact_level	Norm(Impact_level)	Float64	Float64
rownames	Deleted	Repeated Index	-	Object	Deleted

2. References

[1]. T. CALLAS and T. PRINE, "Porsche Airbag Systems | Issue 179 | Excellence | the Magazine about Porsche," Excellence, 2009. https://www.excellence-mag.com/issues/179/articles/porsche-airbag-systems

[2]. M. C. Meyer, T. Finney, and C. M. Farmer, "Who Wants Airbags?," Colorado State University, 2005. https://www.stat.colostate.edu/~meyer/airbags.htm

[3]. Arel-Bundock V (2023). Rdatasets: A collection of datasets originally distributed in various R packages. R package version 1.0.0, Retrieved from https://vincentarelbundock.github.io/Rdatasets/

[4]. "WHAT IS NASS?" [Online]. Available: https://www.nhtsa.gov/sites/nhtsa.gov/files/nassbrochure.pdf

[5]. "National Automotive Sampling System - Crashworthiness Data System (NASS-CDS) - NASS-CDS (multiyear) | USDOT Open Data," datahub.transportation.gov, 1989. https://datahub.transportation.gov/Automobiles/National-Automotive-Sampling-System-Crashworthines/xrgf-q6dn/about_data

[6]. "National Automotive Sampling System - Crashworthiness Data System (NASS-CDS) - NASS-CDS (multiyear)," Data.gov, Oct. 29, 2021. https://catalog.data.gov/dataset/national-automotive-sampling-system-crashworthiness-data-system-nass-cds-nass-cds-multiyea

[7]. Arel-Bundock V, "R: Airbag and other influences on accident fatalities," vincentarelbundock.github.io. https://vincentarelbundock.github.io/Rdatasets/doc/DAAG/nassCDS.html

B. NBA 3-Point Shot Outcome (Report 2) Appendix

1. Metadata and data dictionary:

Metadata		
Name	shot_logs	
Source	https://www.kaggle.com/datasets/dansbecker/nba-shot-logs/data	
File format	csv	
File size	16.4 MB	
Shape	21 columns; 128,069 rows	
Data Dictionary		
Attribute	**Data Type**	**Description**
GAME_ID	int64	Unique numerical identifier for all matches
MATCHUP	object	Unique string identifier for all matches; includes respective date and teams
LOCATION	object	String (‘A’ or ‘H’) specifying whether shooting player is on Home or Away team
W	object	String (‘W’ or ‘L’) specifying whether shooter is on Winning or Losing team
FINAL_MARGIN	int64	Integer corresponding to final point differential
SHOT_NUMBER	int64	Integer representing the shooting player’s total shot count in the game
PERIOD	int64	Integer specifying the quarter of the game in which the shot was taken
GAME_CLOCK	object	String specifying the time left in the quarter in which the shot was taken
SHOT_CLOCK	float64	Float representing the seconds remaining for the offensive team to shoot the ball in the current possession
DRIBBLES	int64	Integer count of the number of dribbles made by the shooting player prior to the shot
TOUCH_TIME	float64	Float representing the seconds that the shooter had possession prior to the shot
SHOT_DIST	float64	Distance (in feet) from the rim that the shot was taken
PTS_TYPE	int64	Integer (2 or 3) specifying the type of shot
SHOT_RESULT	object	String (‘Made’ or ‘Missed’) specifying whether the shot went in the basket
CLOSEST_DEFENDER	object	Name (Last name, First name) of the closest defending player
CLOSEST_DEFENDER_PLAYER_ID	int64	Unique numerical ID of the closest defending player
CLOSE_DEF_DIST	float64	Distance (in feet) between the shooter and closest defending player
FGM	int64	Integer specifying whether the shot was made (1) or missed (0)
PTS	int64	Integer specifying the points resulting from the shot (0, 2 or 3)
player_name	object	Name (first last) of the shooting player
player_id	int64	Unique numerical ID of the shooter

C. Predict likelihood of Term Deposit (Report 3) Appendix

Metadata

Metadata	Information
Name	Bank-full
File format	Csv
File Size	4.6 MB
Dataset Characteristics	Multivariate
Subject Area	Business
Language	English

Data Dictionary

Variable	Pre-processed Variable	Data type	New Data Type	Description
age	age	int64	int64	Age of the client
job	job	object	object	The occupation of the client
marital	marital	object	object	Clients' marital status
education	education	object	object	Educational level of the client
default	default	object	int64	Default credit history
balance	balance	int64	int64	Average yearly balance
housing	housing	object	int64	Housing loan record
loan	loan	int64	Int64	Personal record
contact	deleted	object	deleted	Type of communication
day	date	int64	object	Last contact day of the week
month	date	object	object	Last contact month of the year
duration	duration	int64	int64	The duration of contact
campaign	campaign	int64	int64	Number of contacts within the campaign to the same client
pdays	deleted	int64	deleted	Number of days passed after last contact
previous	previous	int64	int64	Number of contacts before the campaign
poutcome	poutcome	object	deleted	Outcome of previous marketing campaign
y	Term_deposit	object	int64	Do the client subscribed a term deposit