Week 1

This week was about the introduction to ADS2002. During the studio, we had the opportunity to explore various interesting topics through a video briefing delivered by respected mentors.

The topics that were portrayed were Catheter Placement, Monash Solar Farms, and Social Media.

One of the reasons I was drawn towards Social Media data as our first preference was because through all the semesters I have done Applied Data science, this topic seemed to be unique and interesting.

This familiarity with Football gave us an advantage, as we were already acquainted with the available resources and had a basic understanding of the necessary methodologies. On the other hand, while Catheter Placement intrigued us, we realized that we weren't as familiar with the terminology and intricacies of the dataset.

Consequently, we ranked Monash Solar Farm and Catheter Placement as our second and third preferences, respectively. Although these topics were enticing, based on prior knowledge of the ADS2002 group project, we were lured toward the Social Media dataset.

Week 2

After expressing my group project topic preferences, we were assigned to our respective groups based on our choices. Fortunately, I was delighted to be assigned the topic of Social Media, which happened to be my first preference. With our focus set on Social Media, we began exploring the available data related to this subject.

The Social Media dataset consisted of 3 different types of Social Media (Facebook, Instagram, and Twitter). These datasets included terms like comments, shares most liked, and more. Analyzing these terms, we understood their importance in weather prediction and forecasting. Dew point temperature, for example, indicates the temperature at which air becomes saturated with moisture, potentially leading to the formation of dew or fog. Apparent temperature, on the other hand, takes into account factors such as humidity, wind speed, and radiation to determine how the weather feels to humans. Relative humidity provides insights into the amount of moisture present in the air relative to its capacity. By familiarising ourselves with these key terms and understanding their relevance, we laid the foundation for our Melbourne weather prediction project.

This initial exploration allowed us to gain a better understanding of the dataset and the specific variables that would play a crucial role in our analysis and modeling.

Week 3

During Week 3, our group initiated the crucial step of data wrangling. This crucial step aimed to transform our raw, messy Social media datasets into a more structured and usable form. Among the initial challenges we encountered, one of the most prominent was dealing with duplicate values scattered throughout the datasets. These duplicates could potentially skew our analysis results, so it was imperative to address them.

However, duplicates were not the only issue we faced. The datasets were riddled with undefined and missing values, further complicating our data preparation efforts. To ensure the accuracy and reliability of our subsequent analyses, it was imperative to handle these missing values appropriately. This involved techniques such as imputation or exclusion, depending on the specific circumstances of each case.

Perhaps the most significant challenge during this data wrangling phase was the diversity of languages in the text data, especially within user comments. With social media being a global platform, comments came from individuals in different countries, each expressing themselves in their native language. To bridge this language barrier, we contemplated the implementation of a machine-learning algorithm capable of translating these comments accurately. However, we recognized that before delving into such advanced techniques, it was imperative to thoroughly clean and preprocess the data using established data-wrangling techniques. This approach would not only ensure the accuracy of our translation efforts but also lay a solid foundation for the subsequent stages of our data analysis project.

Week 4

In the current week, our focus remained on the critical task of data wrangling. Building on our efforts from the previous week, where we successfully addressed duplicate values in the datasets, our main objective was now to eliminate undefined values. This involved a meticulous process of identifying and removing any data points with missing or undefined values, ensuring that our dataset was clean and reliable for subsequent analysis.

This dataset was related to the FIFA 2022 World Cup, so it was quite imperative that we ensured the timeline matched our data. The timeline was Dates e.g. (1-19th Nov, round dates, 18-30 dec), which refers to Pre-tournament, each round, post-tournament. For example, below is the sample code for the Twitter Dataset.

twitter['Date'] = pd.to\_datetime(twitter['Date']) pre\_tournament = twitter[twitter['Date'] < '2022-11-20'] tournament = twitter[(twitter['Date'] >= '2022-11-20') & (twitter['Date'] <= '2022-12-18')] post\_tournament = twitter[twitter['Date'] > '2022-12-18'] pre\_tournament = pd.DataFrame(pre\_tournament) tournament = pd.DataFrame(tournament) post\_tournament = pd.DataFrame(post\_tournament)

It was also important to come up with a potential research question. We thought of analyzing “Analysing the evolution of trending topics during the 2022 FIFA World Cup.”. We could achieve this by observing and filtering the top likes and comments from each Social Media dataset.

Week 5

This week, we decided to move on to data visualization. We wanted to explore the actual interactions vs. the content discussed on Social Media. However, before we could do this we still had to figure out how to translate languages like Spanish, Arabic, French, etc. into English.

Later, we found a breakthrough when we imported the NLTK pack in python. Which is the Natural Language Toolkit. This is the code we used to function the NLTK pack.

import nltk from nltk.corpus import stopwords from nltk.tokenize import word\_tokenize

nltk.download('stopwords') nltk.download('punkt') stop\_words = set(stopwords.words('english'))

df['Variable'] = df['Variable'].apply(lambda text: ' '.join([word for word in word\_tokenize(text) if word.lower() not in stop\_words]))

* import nltk: It imports the nltk library, which provides various tools for natural language processing.
* from nltk.corpus import stopwords: This line imports a list of common stopwords (i.e., common words like "the," "and," "is," etc.) from the nltk library. These words are often removed from text data because they don't usually provide meaningful information.
* from nltk.tokenize import word\_tokenize: This line imports a function called word\_tokenize, which is used to split text into individual words or tokens.
* nltk.download('stopwords') and nltk.download('punkt'): These lines download necessary data files and resources for the nltk library related to stopwords and tokenization.
* stop\_words = set(stopwords.words('english')): Here, it creates a set of stopwords specifically for the English language. This set will be used later to filter out these common words from text.

This allowed us to handle data wrangling, cleaning and analyzing more effectively.

Week 6

In this week, we decided to get a headstart on the Project report. I was assigned to work on the Background and Data Wrangling section of the report. In addition to these sections, we also discussed the importance of expanding our research questions. Research questions drive the entire project, shaping our objectives and guiding our analysis. By gathering more research questions, we aim to explore the subject matter more comprehensively and deeply. These additional questions will not only enhance the quality of our research but also allow for a more thorough investigation, potentially leading to valuable insights and findings.

Overall, this week's focus on the project report and the expansion of our research questions demonstrates our commitment to producing a well-rounded and impactful research project. We're determined to maintain this momentum and ensure that our project's outcomes are well-documented, rigorous, and insightful.

* What type of content specifically generates the most interest in terms of engagement? (news, WC updates, other)
* Is engagement dependent on the type of user (news outlet, influencer, common user, etc.)
* Which hashtags are the most used in posts? Does using certain hashtags result in higher engagement?
* Does content with positive sentiment generate more engagement than those with negative?
* How do the trending topics evolve throughout the tournament’s progression? Are there specific periods where certain controversies are most highlighted?
* How the trending players changed throughout the course of the World Cup
* How different players’ names are associated with different reactions (Facebook)

Week 7

Unfortunately, I was not able to attend the studio this week as I was not feeling well. However, I managed to update my group on the progress made with the report. The background section and the data wrangling sections were acknowledged and finalized. We have Messenger as our Communication platform, and that’s where we’ve been communicating our doubts or updates. We plan to meet next Tuesday to collaborate further on the Report and project.