**Directions**

1. If you haven’t done so already, [download your Achievement 6 project brief (.pdf)](https://images.careerfoundry.com/public/courses/data-immersion/A6/Data_Immersion_A6_Project_Brief.pdf).
2. The data you use for your project will need to meet certain criteria as defined in the brief. Read through the data requirements now to be sure the data you choose is appropriate.
3. **Source your data.** Use the requirements (and your own interests) to source an open data set from the web. We introduced you to several sources in this Exercise but feel free to look elsewhere, as well.
4. Create a new document to detail your project information.
5. Create a “Data Source” section in your project document and provide the following information:
   * A summary of your data source. We recommend you revisit [Exercise 1.4: Sourcing the Right Data](https://careerfoundry.com/en/steps/sourcing-the-right-data) for a recap on what to include in your summary.
   * An explanation for why you’ve chosen this data set.
6. **Clean your data.** Conduct some basic data cleaning and consistency checks in Jupyter to ensure your data is ready for further analysis.
7. **Understand your data.** Develop a basic understanding of your data set by reviewing the variables and performing basic descriptive statistical analysis. You might want to make a data profile similar to what you did in Achievement 1.
8. **Consider limitations and ethics.** Outline any limitations and ethical considerations presented by the content of your data, its source, and/or how it was collected.
9. Include the results of steps 6 to 8 in a second section of your project document. This second section can be titled something like “Data Profile.”
10. **Define questions to explore.** In a third section of your project document, define a list of questions to explore with your analysis. As mentioned in the Exercise, you may want to revisit [Exercise 1.2: Starting with Requirements](https://careerfoundry.com/en/course/data-immersion/exercise/starting-with-requirements) for a recap on writing good questions.
11. Submit your project document and Jupyter notebook to your tutor for review.

The data I have selected comes from this source:

<https://www.kaggle.com/datasets/bhuvaneshprasad/test-matches-dataset-1877-2023?select=test_Matches_Data.csv>

For this project it is vital that my data:

|  |  |
| --- | --- |
| **Requirement** | **Selected Data** |
| Has geographic data | Contains the country of both teams in each match  Contains the country, city and stadium of each match |
| Has at least 3 continuous variables | Scores for all four innings  Match result (in numeric terms) |
| Has at least 3 categorical variables | Team names  Toss winner  Toss decision  Match result – who won / draw / tie |
| Has at least 2,000 rows | 2.537 rows |
| Non-anonymised columns | All well-labelled |
| More recent data | From 1877 up to 2024 |
| Has a time dependent variable | Date of each match |

**Data Sourcing**

This is an external data source.

I have sourced it from Kaggle. I don’t know the accuracy of the data, but have no reason to presume it is inaccurate. It has been used, and commented on by at least two other analysts.

Using another source – CricInfo – I was able check the number of matches recorded is accurate. I will conduct appropriate spot checks on data as I proceed.

**Data Collection**

Data is collected at each match and widely recorded publicly, in media and ultimately in records like Wisden or CricInfo.

Any of these individual pieces of data could be verified using these sources.

**Data Contents**

This data contains a full record of the teams, score, result and venues in each match – as well as data that I am unlikely to need on the p[layers and officials involved.

It links to other data sets that I would be able to draw on – player data and more detailed statistics for each match

**Data Relevance**

As part of this project, I want to see to see how the balance of results has changed over time, how teams from different geographies have performed compared to one another, and ultimately if Test Match cricket has become more “exciting”.

This data will enable me to do all of that.

**Data Cleaning**

Using Python, I performed the following checks:

* Reviewing each column
* Making that view into a .csv
* Checking for distinct values
* Dropping columns
* Renaming columns
* Looking for duplicate rows

And this has produced: “test\_matches\_data\_Clean” with 14 columns removed and 21 columns renamed

Performing a visual review in Excel, it was quickly apparent that the 19th century test matches has a different date format. This is only 64 test matches, and will not greatly affect my findings, so I have removed them.

It was also apparent that there was only one value of “Format” = TEST” – so have removed this column too.

**Data Consistency**

Importing this back intot Python. I looked at the basic descriptive stats of the continuous variables, and I found nothing unusual.

Downloaded as a .cvs called “df\_stats”

I also looked at unique values, and investigated an anomaly, which was easily resolved

I created a subset one team’s matches – which turned out be only one match

So, I made no further changes to the data.

**Data Limitations**

We don’t have data before 1900 or after 2024, but I don’t consider this a risk to this analysis.

We have all the data we do need – which appears to be a complete record – to conduct interesting and useful analysis.

**Data Ethics**

I do not see that there are any ethical considerations from suing this data – it’s very public and freely available

**Data Bias**

There is Montgomery bias in the data, it is a record of the events. I must be mindful of my bias and preconceptions when analysing this data.

**Questions to explore**

Ultimately, I want to find out if test cricket has become more “exciting”.

That will be determined by:

* Are there more results?
* Does more - more runs, more wickets – happen in a shorter period of time?
* Are there more errors (extras)?

I would like to look at the patterns of these over time.

I would also like to see if home advantage and geography has an affect on results, and if these have changed over time.