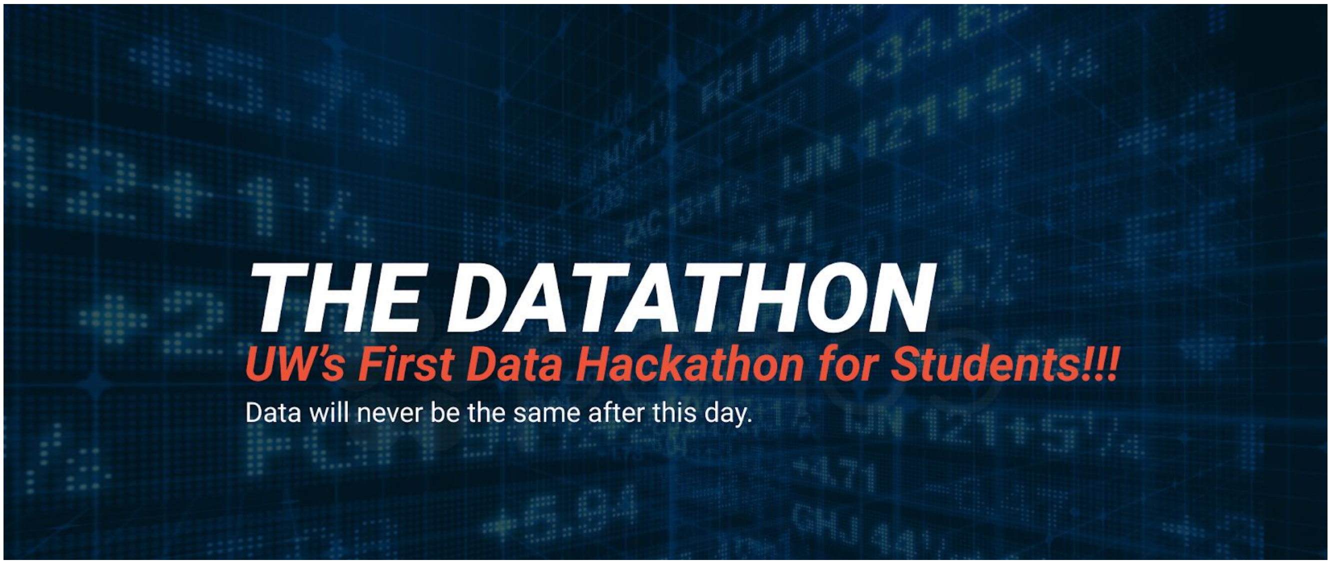
THE DATATHON
UW's First Data Hackathon for Students!!!
Data will never be the same after this day.

Hi,

Welcome to the Dubstech Datathon. The University of Washington's first homegrown data science hackathon. This document entails the rules of the competition, the prompts and the awards for today.

Rules:

- You can participate in a team of maximum 5 people (recommended) or individually.
- Each team can only pursue one prompt and make one submission.
- You are allowed to use external datasets provided that you put a link to them.
- You are not allowed to submit projects completed outside of the event as your submission.

Prizes (all our prompts qualify for these prizes):

- Best Data Analysis and Insights | 1st place, 2nd place, 3rd place
- Best Data Visualization / Infographic / Dashboard | 1st Place, 2nd Place, 3rd Place
- Best Recommendation/Prediction Model | 1st place, 2nd place, 3rd place
- Best Metric | 1st Place, 2nd place, 3rd place

A few things to note before you start:

- You only have 3 hours to work on your project. Select a prompt whose domain you are familiar with or which interests you and stick with it. We also recommend that you leverage your existing skills to extract insights and not spend too much time learning new tools.

- If you are writing a report, we strongly encourage you to write your notes and descriptions as you progress through the data with your code

- If you are making a complex interactive visualization, we strongly encourage you to invest time into making only 1 of these. Only if your visualizations are more basic, do we suggest you make more than 1 visualization. Do not forget to highlight the insights from your visualization.

Schedule

We will have an early check-in Denny Hall to help you form your team and read the prompts carefully.

4 PM - 5 PM - Early Check-In + Icebreaker session + Team Formation
5 PM - 8 PM - Work on your prompts and submit your work here: https://goo.gl/forms/jepuDk7ktpj4OVJ82
8 PM - 9 PM - Project Showcase and Judging
9 PM - 9:15 PM - Results & Prizes

Prompts

scroll down to read more

Students who attended the Citadel Datathon cannot attempt this prompt

Improve Airbnb in Seattle & Boston

Seattle and Boston are two of the biggest business and innovation hubs in the country, attracting heavy traffic from tourists and professionals alike. The cities draw people from all walks of life ranging from computer scientists to business owners to startup specialists to tourist groups to college freshmen and their anxious freshman. Airbnb senses an opportunity to improve their rental programmes in these cities and would like to hear your suggestions on the same.

Airbnb wants you to conduct a study on how they can improve their current rental programmes for tourists and visiting professionals in either or both of these cities.

Airbnb Seattle & Boston Datasets: https://goo.gl/jcHuwG
Sourced from: https://www.kaggle.com/airbnb/seattle, https://www.kaggle.com/airbnb/boston

Listings: details about each rental property available to customers
Calendar: when and what cost is each listing available
Reviews: reviews left by customers

Additional Datasets: Demographics, Econ State, Real Estate Prices, Venues

Possible Questions to Explore/Ideas:
These questions are for your guidance. We encourage you to look at the data and make questions of your own.

- Is there an upward trend in new Airbnb listings and total Airbnb visitors to Seattle?
- What is the expected demand and supply for Airbnb rental properties required for the next 3 years?
- Could prices/amenities be improved to help increase customers for a property?

Deliverable Requirements

Get Datasets Here: https://goo.gl/jcHuwG

The manner in which you provide your insights and recommendations is up to you. We recommend that you focus on providing a maximum of 3 insights/recommendations. You will be graded on quality and not quantity.

It can be an analytical report or a prediction model or data visualization or infographic or dashboard.

What to Submit? (Submissions expected by 8 pm)

- Report or Tool in a universally accessible format - Website Link / Notebook / PDF / PPT
- Source Code to generate the report

Celebrating 120 years of the Olympics

Held every four years, the Olympic Games are considered the world's foremost sports competition with athletes from more than 200 nations participating in a variety of sporting competitions. Being the oldest and the grandest sporting event, a large amount of data has been acquired from the games' history.

As part of their 120 years celebration, the Olympic committee wishes that you publish a mini case study that highlights significant insights and makes recommendations for future events.

Datasets: https://goo.gl/dcxmRD

This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. Note that the Winter and Summer Games were held in the same year up until 1992. After that, the Winter games occurred separately occurring every four years starting with 1994.

- **Athlete Events**: Data on each athlete entry for each competition for each year
- **NOC Regions:** The National Olympic Committee Code and it's region (for map visualizations)

Columns Available: Name, Sex, Age, Height, Weight, Team, NOC (country code), Games, Year, Season, City, Sport, Event, Medal, Region Name, Notes

Possible Questions to Explore/Ideas:

These questions are for your guidance. We encourage you to look at the data and make questions of your own.

- How has the representation of males and females evolved over time?
- Does location affect the performance of competitors? (i.e. "home field advantage")
- What combinations of height and weight show the best results in different sports?
- Idea: Develop a metric to evaluate the most exciting Olympic event & country progress

---

Deliverable Requirements

Get Datasets Here: https://goo.gl/dcxmRD

The manner in which you provide your insights and recommendations is up to you. We recommend that you focus on providing a maximum of 3 insights/recommendations. You will be graded on quality and not quantity.

It can be an analytical report or a prediction model or data visualization or infographic or dashboard.

What to Submit? (Submissions expected by 8 pm)

- Report or Tool in a universally accessible format - Website Link / Notebook / PDF / PPT
- Source Code to generate the report

Metrics for European Soccer Leagues

The last 10 years of European soccer has been extremely exciting. With the transfer records being broken on multiple occasions to underdogs showcasing extremely high skill to teams showcasing extreme dominance on the field. As part of their efforts to assist clubs and pundits, Optasports and UEFA are currently building a set of metrics to be used for player, team and league evaluations.

Your challenge is to investigate the available data and develop a metric or metrics and present their application through a results report which demonstrates how it is used.

Datasets:

Top 5 European Leagues: Statistics and Betting odds for the last 10 years of league matches
Fifa Ratings: Statistics of each player for the years 2017, 2018, and 2019
European Database: Detailed records of each player / team / league | Learn more on how to use
Note: Some of these datasets will have more columns with time.

Possible Questions to Explore/Ideas

- What are the most exciting players/teams/leagues in Europe?
- What are the trends in the players of teams and leagues?
- Can we predict the growth of a player based on the league and team they are a part of?
- Which league is the most suitable for particular type and age group of player?

We recommend reading the column names before developing your metrics.

Deliverable Requirements

Get Datasets Here: https://goo.gl/P7KHwD

The manner in which you provide your metrics and results is up to you. We recommend that you focus on providing a maximum of 3 metrics. These can be submitted through an analytical report or a prediction model or data visualization or infographic or dashboard. You will be graded on quality and not quantity.

You must include an explanation of your metric, how it came to be, and when it should be used.

What to Submit? (Submissions expected by 8 pm)

- Report or Tool in a universally accessible format - Website Link / Notebook / PDF / PPT
- Source Code to generate the report

The Age of Kickstarter

Kickstarter is a funding platform where creators can share and gather interest in a particular creative project they'd like to launch. It's entirely driven by crowdfunding, where the general public and their money is what sends these projects into production. Every project is independently crafted while friends, fans and total strangers offer to fund them in return for rewards or the finished product itself.

Kickstarter wants you to create a study that provides significant insights and helps expose them to and projects and categories they should pay attention to for the upcoming year.

Dataset: https://goo.gl/3qASjX

This dataset contains information about over 300,000 Kickstarter projects, with information such as category, goals, and pledges.

Column Definitions

- *usd_pledged:* conversion in US dollars of the pledged column
- *usd pledge real:* conversion in US dollars of the pledged column
- *usd goal real:* conversion in US dollars of the goal column

Possible Questions to Explore/Ideas

- Is there a correlation between the goal of the project and its success?
- How can this data help individuals and startups that wish to launch their idea on Kickstarter?
- Are there certain types of media more prone to success on the platform?
- What is the forecast of new projects and funders by category for the upcoming year?

Deliverable Requirements

Get Datasets Here: https://goo.gl/3qASjX

The manner in which you provide your insights and recommendations is up to you. We recommend that you focus on providing a maximum of 3 insights/recommendations. You will be graded on quality and not quantity.

It can be an analytical report or a prediction model or data visualization or infographic or dashboard.

What to Submit? (Submissions expected by 8 pm)

- Report or Tool in a universally accessible format - Website Link / Notebook / PDF / PPT
- Source Code to generate the report

Prior knowledge of Sentiment Analysis is highly recommended

Sentiment Analysis of the WorldCup 2018

The Fifa World Cup 2018, the most prestigious association football tournament, as well as the most widely viewed and followed sporting event in the world, was one of the Top Trending topics frequently on Twitter while ongoing.

Twitter wants a sentiment analysis study which investigates and highlights the different emotions people experienced during the world cup between the Round of 16 and Final.

Dataset: https://www.kaggle.com/rgupta09/world-cup-2018-tweets/home

This dataset contains a random collection of 530k tweets starting from the Round of 16 till the World Cup Final that took place on 15 July 2018 & was won by France.

Possible Questions/Ideas to Explore:

- Common Patterns and Trends in sentiments expressed for each match
- Visualization of how sentiments changed between the round of 16 and final
- What kind of sentiments get the most retweets?

Sentiment Analysis of Charlottesville Rally

Robert E. Lee was a US Army general who defected to the Confederacy during the American Civil War and was considered to be one of their best military leaders. His statue in Charlottesville is slated to be removed. While many Americans support the move, believing the main purpose of the Confederacy was to defend the institution of slavery, many others do not share this view and have not taken its planned removal lightly.

The Unite the Right rally (Charlottesville rally) was a white supremacist rally in protest of the removal of the statue that occurred in Charlottesville, Virginia, from August 11 to 12, 2017. Protesters were members of the far-right and included self-identified members of the alt-right, neo-Confederates neo-fascists, white nationalists, neo-Nazis, and various militias. Tragically, one of the counter-protestors to this rally--Heather Heyer--was killed and many others injured after a man intentionally rammed his car into them.

This dataset below captures the discussion--and copious amounts of anger--revolving this event. It contains a random sample of 50,000 tweets from between August 15 to August 16, after the event.

Dataset: https://www.kaggle.com/vincela9/charlottesville-on-twitter

Twitter wants a sentiment analysis study which investigates and highlights the patterns in the emotions people portrayed after this rally.

Possible Questions/Ideas to Explore:

- How does proximity to the event effect tweet word choice and sentiment?
- Does follower count correspond to how quickly they started tweeting about the event?
- Does word choice such as political stances in user description correspond to tweet text?

Deliverable Requirements

The manner in which you provide your insights and results is up to you. These can be submitted through an analytical report or a prediction model or data visualization or infographic or dashboard. You will be graded on quality and not quantity.

You must include an explanation of what sentiment analysis method you used and provide a dataset which shows the words classified by emotion.

What to Submit? (Submissions expected by 8 pm)

- Report or Tool in a universally accessible format - Website Link / Notebook / PDF / PPT
- Source Code to generate the report