


May 16 – May 17 | register now on
bit.ly/Datathon20

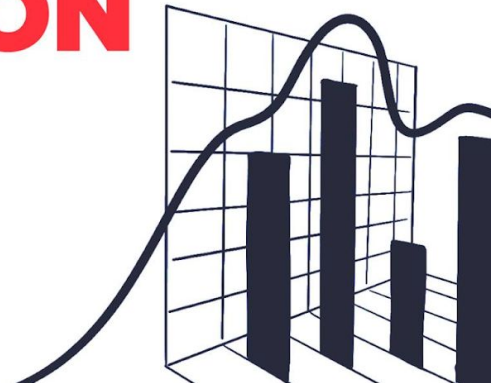
Sponsored by **pathrise**

DATATHON

the online data hackathon

Presented by

DUBSTECH x Machine Learning Society x Applied Analytics x Blockchain Society x  Information School
UNIVERSITY OF WASHINGTON



Hello,

Welcome to the 3rd DubsTech x Machine Learning Society x Applied Analytics x BlockChain Society Datathon, the University of Washington's first homegrown data science hackathon co-powered by Pathrise. This document entails the rules, prompts and the awards for the 3rd Datathon 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 | 1st Place, 2nd Place, 3rd Place
- Best Data Visualization / Infographic / Dashboard | 1st Place, 2nd Place, 3rd Place
- Best Machine Learning Model | 1st Place, 2nd Place, 3rd Place
- Best Insight | 1st Place, 2nd Place, 3rd Place

A few things to note before you start:

- You do not need to pursue each of the prizes
- Select a prompt whose domain you are familiar with or which interests you
- Leverage your entire team's skills to extract insights (ask each other about your strengths)
- We strongly encourage you to write notes and descriptions in the report as you write your code

Schedule

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

Join All Meetings via washington.zoom.us/j/99745378938

Day 1

10:00 am: Opening Ceremony
10:30 am: Team Building Event
11:00 am: Workshop 1 - Data Cleaning with Excel
12:15 pm: Workshop 2 - Data Visualization with Power BI
1:30 pm: Workshop 3 - Data Visualization with Tableau
3:30 pm: Workshop 4 - Machine Learning with Azure ML Studio
5:00 pm: Happy hour + Q & A

Day 2

10:00 am: Opening notes
11:00 am: Office hours + Help session
2:30 pm: Submissions Due
3:00 pm: Judging Starts
4:00 pm: Happy Hour + Project Showcase
7:00 pm: Results + Closing Ceremony

COVID-19 in India

Data Repository: github.com/zcolah/COVID_19_Datathon

Task Category: Data Analysis, Machine Learning Modelling, Data Visualization

The COVID-19 pandemic which has spread around the world in less than 6 months, with more than [4.56 million cases](#) reported in more than 188 countries and territories, resulting in more than 308,000 deaths. So far the most affected countries have been China, Iran, Italy, Spain, USA, and UK.

Due to the large risk of the spread of the virus that could spell upon the vast population of India, It's government placed all 1.3 billion citizens under a nationwide lockdown. The lockdown went into effect on March 25 and was extended until at least May 17. State borders have been sealed, schools and religious sites were closed, and officials are urging people to stay inside as much as possible.

Despite these efforts, India's cases have been on the rise, with no indication of it flattening or decreasing any time soon. With many fearing the worst is yet to come, the government has handed you a set of datasets (see next page) so that they can get an accurate sense of the issue at hand.

Your task is to answer one or more of the following questions listed below:

- **Estimate at which point can one expect to see the peak in number of COVID-19 cases in:**
 - India as a whole
 - The following states: Maharashtra, Delhi, Tamil Nadu, Gujarat
 - The following districts: Mumbai, Delhi, Ahmedabad, Chennai
- **By which date will the states and districts listed above have less than 100 COVID-19 cases?**
- **Based on the data we've provided - what would be the number of cases in the aforementioned states and districts by June 30th?**
- **Estimate rate at which the virus has spread across the country?**
- **How has the growth and the spread of the virus looked across the country since the first case?**
- **Make meaningful visualizations that can effectively display the current issue at hand.**

Submission Requirements

Submission Format: The manner in which you present analysis and answers should be in a universally acceptable format (.doc, .pdf, .md, url) along with a link to the Github repo. If you make a machine learning model, please include your accuracy and margin of error for your model.

We recommend that you focus on answering a maximum of 3 questions. Note that if you are making a machine learning model you do not need to try and make it predict the values for all states or districts mentioned. You will be graded on quality, not quantity.

COVID Datasets	
Dataset Name	Dataset Description
district_daily_may_15.csv district_daily_may_15.json	A CSV file with total confirmed, active, deceased, recovered, by date for each district in India since 21st April, 2020. Also available in JSON form.
patient_city_district_wise_data_may_5.csv	A CSV file with data of each patient and their respective city and district data until 5th May, 2020.
patient_city_district_wise_data_may_5_date._formatted.csv	A CSV file with the same data but with date column formatted in a universal readable format.
state_wise_daily_delhi_gujarat_maharashtra_tamil_nadu_may_15_2020 - state_wise_daily_mh_gj_dl_tn.csv	A CSV file with sorted date column, status of patients(recovered, deceased, confirmed) and statewise data(from Delhi, Gujarat, Tamil Nadu and Maharashtra) about the number of patients with that recovery status on the given date.
Additional Datasets	
HospitalBedsIndia.csv	A CSV file with state wise data about the number of healthcare facilities and beds for that state.
ICMRTestingLabs.csv	A CSV file with data about testing facilities across the country, their addresses, city, state and type of facility(Government lab, collection site etc.)
StatewiseTestingDetails.csv	A CSV file with state wise data of total samples collected, number of positive and negative samples and the date.
district_population_india_census2011.csv	A CSV file with census data at the district level for the year 2011.
state_population_india_census2011.csv	A CSV file with census data at the state level for the year 2011.
zones.csv	A CSV file with information about the latest state of each zone in terms of cases(Red - bad condition, lot of cases; Orange- fewer cases than red; Green- Less number of cases not in a very bad condition) with district and state information.



Grocery Retail Sales Data

Dataset: drive.google.com/file/d/1FGjwCK46XSZP2bYLbzTnMb97jC9DIIEu/view?usp=sharing

Task Category: Data Analysis, Machine Learning Modelling, Data Visualization

A small grocery store in Australia has been registering every transaction happening in its store from the year 2016 to the year 2019. Now that they have sold over 650,000 items from their store, they have decided to focus on using this data to help themselves make better business decisions.

Before you start with this prompt we recommend [reading this report as a reference](#).

Your task is to answer one or more of the following questions listed below:

- **Predict the total amount of sales for the store, for each category, and for each item in the next 3 months. (You can define your own metrics if you wish to.)**
- **What trends do you notice for the store with respect to transactions?**
- **What trends do you notice for the store with respect to time?**
- **What trends do you notice with respect to the categories & items listed?**
- **Are there any categories or items the store should immediately focus on?**
- **Given Inventory levels, price, rate of purchase etc., what are some (1 to 2) products that can be considered for a price reduction?**
- **Give a sales analysis report answering fundamental questions that the store owner wants to know:**
 - The highest selling products by month and category
 - The least selling products by month and category
 - The most profitable month by sales
 - What are the most efficient ways that the store can reduce losses
 - Any other question you can think of!

Submission Requirements

Submission Format: The manner in which you present analysis and answers should be in a universally acceptable format (.doc, .pdf, .md, url) along with a link to the Github repo. If you make a machine learning model, please include your accuracy and margin of error for your model.

We recommend that you focus on answering a maximum of 3 tasks. You will be graded on quality, not quantity.

Judging Rubric

Data Analysis

Criteria	Points Available
Explanation of Process used is provided and detailed	5
Quality of Exploration of Data	5
Quality of Metric Created	5
Description of Metric	5
Entities ranked based on metric created	5
Total	25

Machine Learning Modelling

Criteria	Points Available
Explanation of Process used is provided and detailed	5
Quality of Exploration of Data	5
Quality of Model Created (features selected, the margin of error)	5
Description of Model Created	5
Value of the Model created	5
Total	25

Data Visualization

Criteria	Points Available
Choice of Colors & Fonts Used	5
Level of Detail Provided	5
Quality of Presentation	5
Explanation of Process used is provided and detailed	5
Quality of Analysis and Exploration	5
Total	25

Questions

Ask your questions on slack: <https://app.slack.com/client/T0149PRCCC8/C013RN5GN4B/details/info>
and get help from our mentors: Saketh, Zoshua, Shray, Shravan, and Ranjith.