Datahack Welcome Packet

Welcome!

We are super excited to see you here at the Datahack! This document contains lots of important info that will come in handy throughout the day. Make sure to skim it right now and come back to it later as needed.

Schedule

- 8:00am - Breakfast - Taco Deli
- 8:30am - Kick-off
- 9:00am - Get to work!
- 9:30am - Oracle Cloud Workshop (Optional - EER 0.810)
- 12:00pm - Lunch - Hoa Hoa
- 12:30pm - Teza Workshop (Optional - EER 0.810)
- 3:15-4:15pm - Judges' office hours
- 4:30pm - Reports due
- 5:30pm - Winners Announced
- 6:00pm - Event Ends

Check In

The check-in is scheduled for 8:00am, right outside EER room 1.518. Make sure you grab your colored sticker for food.

Food

Meals

We made sure that we have enough food for everyone! You should have received a colored sticker at check in. Line up when your color is called.

Snacks

Feel free to grab snacks from the assorted pile at any point!

Mentors

We will have mentors around all day. Look for someone not furiously trying to get a report done. Feel free to flag us down and we will come over to you to help you out. All mentors will be wearing a nametag.

Datasets

Obtaining the Datasets

We have found some great datasets and cleaned them up for you. We are sending out some limited information about it now. You can access the datasets at https://goo.gl/GFh18Y.

Obtaining the Starter Code

You can access the starter code at https://goo.gl/JpZGXw. Feel free to ask a mentor if you have trouble accessing code from Github.

Accessing the Datasets

Here is a description of the datasets that we will provide you. There will be more info here on the day of the hackathon. **You are allowed to bring in supplementary datasets, but must place emphasis on at least one of the datasets that we provide you.**

Dataset I - Education in India

The theme for today is education. The importance of education is self-evident, even more so in developing countries. Differences in access to quality education lead to drastically changed lives. Education affects all spheres of life, and is the powerhouse of an economy. Will you be the one to mine a world-changing insight out of the following dataset today?

Some recommended articles to provide context about challenges in education and the potential impact of data science include:

Why the World's Biggest School System is Failing Its Pupils
Can Big Data Change a Wicked School Truancy Problem?
'Big Data' Was Supposed to Fix Education. It Didn't.

Here is where we originally got this dataset
https://www.kaggle.com/rajanand/education-in-india

Dataset II - Health in India

Note that this dataset will be great to use in conjunction with the first one. We will provide you granular data about health in India. Perhaps you can find a way to correlate it with the former.

http://debarghyadas.com/writes/nfhs-4/

Dataset III - World Education Dataset

This dataset will contain different education indicators for countries all over the world. Each row describes a certain country's indicator values over a period of time. Note that for each row, some years may be empty, but no row will have zero information for every year.

Column/Row descriptions:

Dimensions: 353,907 rows *x* 61 columns

Metadata:

Column Name	**Description**
Country Name	Name of country. *String*
Country Code	Abbreviated country. *String*
Indicator Name	Statistic that specific row is describing. Contains 3617 unique values. Pay attention to the units defined at the end of each name. *String*
Indicator Code	Period delimited code that is unique to each Indicator Name. *String*
Years (Multiple Columns)	Data collected for given indicator in a given year. Unit will be defined by the Indicator Name. *Float*

Here is where we originally got the dataset:

https://www.kaggle.com/theworldbank/education-statistics

Dataset IV - United States Education, Labor, and Healthcare

This dataset includes national and state-level information. The indicators include **healthcare** metrics such as birth rate and infant mortality rate, **education** metrics such as enrollment and population with a college degree, and various others in **crime,**

income, labor, economy, and welfare.

Here is where we originally got the dataset: https://knoema.com/bjcvbhc/us-regional-dataset.

Submission

You can submit at https://texasdatahack.devpost.com/. Note that most announcements will still be made on Slack and you should look there for the most up to date information.

Report

Please submit a report containing your hypothesis, analysis, and any visuals that you would like to include. This will be the primary document that the judges will use to judge your submission, in addition to your presentation. Please cite any supplementary datasets and/or code that you used in your report. **Note that you must prominently use at least one of the datasets that we provide.** We will feature the winning reports on our website.

Report Format

The following are some **required sections** for the report. Having these sections will help the judges in judging your report. You may add additional sections. Remember, the judges will have limited time to go through the reports. We highly recommend limiting the report to not more than 2 pages (excluding visualizations).

Problem Statement
1-3 sentences about what you are trying to do.
Eg. 1. Can we find predict the quality of wine based on the year the grapes were harvested?
Eg 2. Can we visualize the birth rates and death rates in Utah over time?

Introduction
Introduce the rest of the report and briefly describe the report.

Datasets
List the provided datasets you used and cite any external ones you used too. *You must use at least one one of the provided datasets.*

Technical Analysis/Methods used

Describe the techniques you used for your project

Results
What were the results? Eg. Accuracy of the classifier. If you built on top of something else, how did you improve upon it?

Conclusion
Conclude your report here.

Citations
Please cite any external code (if it forms a non-trivial part of your project) and external datasets you use.

Code (.zip) (optional)

Please submit any code you wrote. This is marked optional, but we may request your code at any time for judging purposes. We will also feature it on our website.

Judging

Judges will be hosting office hours, you may sign up for a 5 minute help slot between 3:15 and 4:15 to get feedback from a judge before submitting your final report at 4:30.

Judges will then score reports and choose winners based off the following criteria:

- Creativity: How much did you differ from the rest of the reports seen? What new things did you try that haven't been done before?
- Social impact: How much were you aware of the implications of your work. How much social good can come from your insights?
- Insights into the data: What did you discover? How well did you get to know the data? What questions did you ask and answer?
- Technical mastery: What skills did you use? How technically complex was your project? How well did you pull it off?

Your report is your chance to showcase your work and tell the judges why you deserve to win!

There will also be a prize awarded for best visualizations.

Mentoring

Physical mentors on site

We will have mentors in the room at all times. Feel free to flag one down and come talk to them at any time. All mentors will be wearing name tags.

Slack (texasdatahack.slack.com)

Feel free to message the mentors on slack. Also feel free to message the `#datasets` channel if you have any questions about the datasets.

Judges' Office Hours

You will be able to sign up for ONE ten minute "office hour" with one judge. The judge will be randomly assigned to you. The office hours will be from 3:15pm to 4:15pm. This is a great opportunity to get some facetime with the judges and incorporate any small changes they suggest for your report. Note that these are optional, so if you would rather be working on your project at that time, that is totally okay. We will post sign up info the day of the hackathon.

Prizes

Thanks to our friends at Oracle, Teza, and OJO labs, who have been extremely generous, we have some awesome prizes!

1st Prize: $1000 in scholarships, $2,500 in Oracle cloud credits
2nd Prize: $750 in scholarships, $1,500 in Oracle cloud credits
Best Visualization: $500 in scholarships, $1000 in Oracle cloud credits

The scholarship money can be traded for any of the following prizes (must add up to be less than or equal to the dollar amount awarded to you):

- **Headphones**
  - Bose Quiet comforts https://goo.gl/WUgFWb ($349)
  - Beats Solo 3 https://goo.gl/jjVLfq ($219)
  - AudioTechnica M50X https://goo.gl/2xhsHz ($149)
- **Home Assistants**
  - Homepod https://goo.gl/uaLa14 ($350)
  - Echo https://goo.gl/amGFBS ($85)
  - Home https://goo.gl/kuHtwy ($129)
- **Tablets**
  - Surfacebook (1st edition) https://goo.gl/V68bey ($855)
  - iPad https://goo.gl/wjiMJW ($450)

- **Smart watches**
    - Apple Watch https://goo.gl/8zFQ36 ($399)
    - Fitbit https://goo.gl/LhbZbJ ($149)
- **E-Readers**
    - Kindle https://goo.gl/RCNWuK ($120)
- **Other**
    - Apple TV 4k https://goo.gl/voabhA ($179)
    - Chromecast https://goo.gl/8zuwQI ($35)

Etiquette

This is a gentle reminder that all the mentors and folks helping out are doing that on a purely voluntary basis. Be nice to them and to your fellow hackers. If you observe any sort of harassment or otherwise unacceptable behaviour, alert a mentor immediately, either in person or through Slack. Bad behavior is grounds for immediate dismissal.