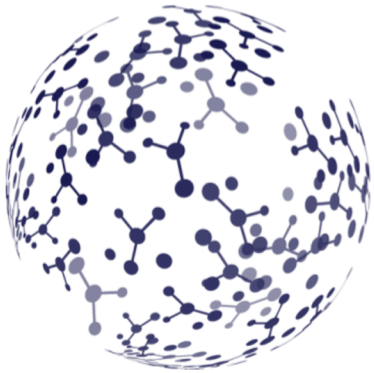


DATASCIENCEGO

HACKATHON

Thanks to our Sponsor!

ORACLE Cloud Infrastructure

DATASCIENCEGO HACKATHON

ORACLE Cloud Infrastructure

DATASCIENCEGO HACKATHON

Introduction

John Peach

- Principal Data Scientist at Oracle
- Data Science Service
- 30+ years of experience in tech
- Orange County R User Group
- https://www.linkedin.com/in/jpeach/

Welcome to the Hackathon!

Goals:

- Practice your skills and learn from others
- Expand your network while collaborating with peers
- Have fun!

General Rules:

- Be respectful and collaborative
- Submit ONLY the work you produce during the Hackathon
- Be honest and ethical. Cheating is easy and will ruin your experience

Agenda (Pacific Time zone)

(NOW) Event Opening: 9:30 AM - 10:00 AM

Working Block #1 & Mentoring Sessions: 10:00 AM - 12:00 PM

Networking Break! 12:00 PM - 1:00 PM

Working Block #2 & Mentoring Sessions: 1:00 PM - 3:00 PM

SUBMISSION OF RESULTS DEADLINE: 3:00 PM

Team Presentations: 3:00 PM - 4:30 PM

Virtual Happy Hour! 4:30 PM - 5:30 PM

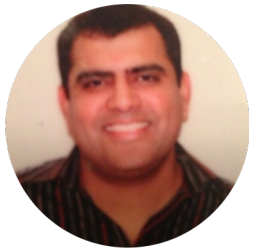
Prize Awarding and Event Closure: 5:30 PM - 6:00 PM

Support Team

Facilitators:

John Peach
Principal Data Scientist at Oracle

Raj Krishnamoorthy
Master Principal Enterprise Cloud Architect at Oracle



Haroon Anwar
Master Principal Architect - Enterprise Cloud Architecture at Oracle

Jiayuan Yang
Staff Solution Engineer at Oracle

(Haree) Srihareendra Bodduluri
Cloud Software Engineer at Oracle

Raja sekhar Reddy
Cloud Software Engineer at Oracle

Support Team

Facilitators:

Anna Fedotova
Product Data Analyst at SuperDataScience

Diane Pesquera
Event Coordinator at DataScienceGO

Ankit Jain
MSBA Program Ambassador |
Graduate Student at UC Irvine

Jordan Sauchuk
Data Scientist at SuperDataScience

Edis Gonuler
Data Science Enthusiast

Communication Tools

ORACLE Cloud Infrastructure | DATASCIENCEGO HACKATHON

Meet and Chat HERE

Share Files, Submit Results and Ask for Support HERE

Download Hackathon Materials HERE

United States Environmental Protection Agency Challenge

ORACLE Cloud Infrastructure

DATASCIENCEGO HACKATHON

- EPA Air Quality monitoring sites
- Identify actionable insights based on the data
  - Best Insights
  - Best Visualization
  - Best Model
- Collaborative team effort
- Most helpful participant across teams
- Create an insightful visualization on the air quality.
- Create a predictive model that will estimate NO3.

The Challenge

3 Different Paths - Align Expectations

- Exploratory Analysis: Understand the data, gather insights and focus on an in-depth analysis of the relationships between the measures, the effects over time or between sites.
- Visualization: Create a visualization that communicates a message about what is happening with the air quality. The visualization can be a figure, set of figures, animation or interactive graphic.
- Model: EPA has identified that the total nitrate (NO3) is a metric that they are interested in studying and predictive model that will allow them to estimate the nitrate values when they are missing

ORACLE Cloud Infrastructure

DATASCIENCEGO HACKATHON

Relevant files

Files

- Air_status.csv – Data for the challenge
- Codes.csv – codes for the air status file
- Data_dictionary.csv – data dictionary
- Site.csv – Site location
- Test.csv - for models
- The Challenge.doc – Challenge description

The Data

SITE_ID	DATEON	DATEOFF	TSO4	TNO3	TNH4	Ca	Mg	Na	K	Cl
CON186	6/17/03	6/24/03	1.646	2.6535	1.074	0.1625	0.044	0.198	0.0609	NA
CON186	6/24/03	7/1/03	1.0356	0.6297	0.4035	0.2331	0.0394	0.1455	0.0444	NA
CON186	7/1/03	7/8/03	1.5335	1.257	0.4836	0.2185	0.0759	0.344	0.1697	NA
CON186	7/8/03	7/15/03	1.7773	0.8323	0.6537	0.2894	0.0541	0.2134	0.085	NA
CON186	7/15/03	7/22/03	2.1637	1.3187	0.7899	0.4331	0.08	0.2679	0.1382	NA
CON186	7/22/03	7/29/03	2.3961	1.8747	1.0025	0.3808	0.0721	0.2498	0.1085	NA
CON186	7/29/03	8/5/03	1.689	0.9653	0.6495	0.2078	0.0383	0.1394	0.0512	NA
CON186	8/5/03	8/12/03	0.9753	0.6864	0.3571	0.2145	0.036	0.1266	0.0401	0.0137
CON186	8/12/03	8/19/03	1.7553	1.8385	0.6672	0.4243	0.0839	0.3391	0.141	0.0143
CON186	8/19/03	8/26/03	1.6582	1.3791	0.5604	0.3728	0.0685	0.2213	0.1159	0.0139
CON186	9/2/03	9/9/03	1.3079	0.951	0.5297	0.2329	0.0421	0.151	0.0532	0.0134
CON186	9/9/03	9/16/03	1.3556	1.3186	0.5847	0.3147	0.0553	0.2038	0.0597	0.0137
CON186	9/16/03	9/23/03	0.8147	0.7078	0.3528	0.3317	0.0455	0.1077	0.0509	0.0131
CON186	9/23/03	9/30/03	1.1388	0.9686	0.4645	0.3119	0.0482	0.1195	0.0488	0.0131
CON186	9/30/03	10/7/03	1.226	0.8245	0.4818	0.2236	0.0328	0.1023	0.0392	0.0132
CON186	10/7/03	10/14/03	1.2985	0.9536	0.5754	0.2748	0.0418	0.1101	0.0466	0.0132
CON186	10/14/03	10/21/03	0.6403	0.3421	0.2585	0.1776	0.0211	0.0485	0.029	0.0133
CON186	10/21/03	10/28/03	0.4887	0.2065	0.1922	0.2103	0.0222	0.0286	0.0537	0.0162
CON186	10/28/03	11/4/03	0.5005	1.5469	0.554	0.1019	0.0242	0.1143	0.0615	0.092
CON186	11/4/03	11/11/03	0.4534	2.7171	0.7852	0.1553	0.0181	0.0532	0.0305	0.0136
CON186	11/11/03	11/18/03	0.3543	0.4171	0.1796	0.0424	0.0098	0.0578	0.0128	0.0139
CON186	11/18/03	11/25/03	0.3248	0.5564	0.1778	0.1	0.0148	0.0509	0.0171	0.0139
CON186	11/25/03	12/2/03	0.3231	0.3933	0.13	0.0973	0.0178	0.074	0.0186	0.0187
CON186	12/2/03	12/9/03	0.259	0.7511	0.2517	0.0847	0.0116	0.0298	0.0153	0.0135
CON186	12/9/03	12/16/03	0.2727	0.7361	0.191	0.0656	0.021	0.0828	0.0151	0.014
CON186	12/16/03	12/23/03	0.0995	0.1531	0.0406	0.0432	0.0095	0.0209	0.0097	0.0136
CON186	12/23/03	12/30/03	0.2173	0.3228	0.1265	0.0306	0.0111	0.0457	0.0118	0.014
CON186	12/30/03	1/6/04	0.2497	0.2404	0.0977	0.0439	0.013	0.058	0.0138	0.0152
CON186	1/6/04	1/13/04	0.2268	0.1052	0.0994	0.0582	0.0063	0.0242	0.0135	0.0137
CON186	1/13/04	1/20/04	0.4061	2.0132	0.6177	0.1311	0.0189	0.0439	0.0265	0.0164

The Data (Glossary)

COLUMN_NAME	UNIT	DATA_TYPE	DESCRIPTION
SITE_ID		CHAR	Site identification code. See the file site.csv for a mapping of site code to site name.
DATEON		DATE	Date the sample collection began, Local Standard Time; YYYY-MM-DD
DATEOFF		DATE	Date the sample collection ended, Local Standard Time; YYYY-MM-DD
TSO4	ug/m^3	NUMBER	Sulfate (SO4) concentration from Teflon filter; ug/m^3.
TNO3	ug/m^3	NUMBER	Nitrate (NO3) concentration from Teflon filter; ug/m^3.
TNH4	ug/m^3	NUMBER	Ammonium (NH4) concentration from Teflon filter; ug/m^3.
Ca	ug/m^3	NUMBER	Calcium (Ca) concentration from Teflon filter; ug/m^3.
Mg	ug/m^3	NUMBER	Magnesium (Mg) concentration from Teflon filter; ug/m^3.
Na	ug/m^3	NUMBER	Sodium (Na) concentration from Teflon filter; ug/m^3.
K	ug/m^3	NUMBER	Potassium (K) concentration from Teflon filter; ug/m^3.
Cl	ug/m^3	NUMBER	Chloride (Cl) concentration from Teflon filter; ug/m^3.
NSO4	ug/m^3	NUMBER	Sulfate (SO4) concentration from Nylon filter; ug/m^3.
NHNO3	ug/m^3	NUMBER	Nitric acid (NO3) concentration from Nylon filter; ug/m^3.
WSO2	ug/m^3	NUMBER	Sulfur dioxide (SO2) concentration from Whatman filter; ug/m^3.
TOTAL_SO2	ug/m^3	NUMBER	Total sulfur dioxide (SO2) concentration calculated from [wso2]+0.667*nso4]; ug/m^3.
TOTAL_NO3	ug/m^3	NUMBER	Total nitrate (NO3) concentration calculated from [tno3]+0.9841*[nhno3]; ug/m^3.
FLOW_VOLUME	m^3	NUMBER	Flow volume; m^3.
VALID_HOURS		NUMBER	Valid hours during sampling period
COMMENT_CODES		CHAR	Comment codes separated by spaces. See codes.csv
STD2LOCAL_CF		NUMBER	Factor used to convert atmospheric concentrations from standard to local conditions
TEMP_SOURCE		CHAR	Source of mean temperature used in conversion factor. See codes.csv
QA_CODE		CHAR	Quality assurance level of the record. (see QAPP for definition of quality assurance levels). See codes.csv
UPDATE_DATE		DATE	Date of last record update; YYYY-MM-DD

Expected Submission and Deadline

Submission is expected to be made by 3:00 PM PT via **Slack**

Expected Submission:

- Submission can be a slide deck or a GitHub repository, including the notebook that implements the full lifecycle of data preparation, source code, model creation and evaluation.
- Excel, OAC, Notebooks, Results Table in CSV format.
- PDF presentation (3 slides max) with your observations, predictions, and conclusions. At the end of the activity, your team will have 3 minutes to present the results.

ORACLE Cloud Infrastructure

DATASCIENCEGO HACKATHON

Judging Criteria

Criteria for Best Model:

- Any team that presents results of their work that were obtained using statistical modeling will be eligible for the Best Model award.
- Any data modeling method can be used, including:
  - Linear regression, logistic regression, and other common statistical models
  - Machine learning models and algorithms such as decisions trees, Random Forest, support vector machines (SVM), and neural networks
- Models will be assessed on SSE.

ORACLE Cloud Infrastructure

DATASCIENCEGO HACKATHON

Judging Criteria

Criteria for Best Insight:

- All teams that present their work will be eligible for the Best Insight award.
- An insight will be considered as piece of “knowledge” gained through the analysis of the hackathon data set. This could include, for example:
  - an interesting conclusion or understanding about the data
  - a recommendation or suggestion about how something could be changed or done differently, supported by data analysis
  - a link or association between different pieces of information in the data, particularly ones that are unexpected or non-intuitive
- Insights will be judged based upon how much practical impact the insight could have on the subject of the hackathon data set, and how well the associated analysis supports it.

Judging Criteria

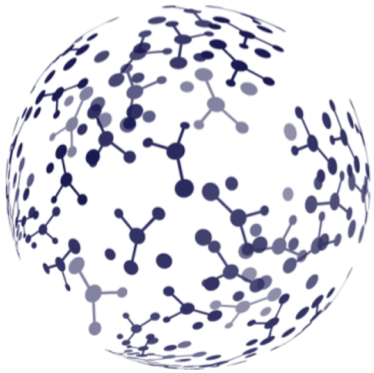
Criteria for Best Visualization:

- Any team that presents a visualization as part of their final submission is eligible for the Best Visualization award.
- Visualizations can be either static graphics (e.g. a plot or 2-D graphic), or an interactive/dynamic visualization (e.g. Oracle Analytics Cloud)
- Visualizations will be judged primarily on their ability to clearly convey an interesting and/or insightful aspect of the hackathon data set. Factors include:
  - how data visualizations are used to communicate interesting aspects of the data
  - the ability for someone not familiar with the data to understand (with the accompanying presentation) the main points of the visualization
  - the effective use of data graphics principles including, for example, the use of statistical summarizations, color, data sub-setting, data highlighting, etc.
  - For static graphics, overly complex visualizations that detract from understanding the points of the visualization should be avoided.
  - for interactive/dynamic graphics, the interactive aspect should be a fundamental component to understanding the points of the visualization. Interactivity for interactivity-sake should be avoided.
  - interactive graphics will not be judged as “better” than static graphics.

Judging Criteria

Criteria for Most Helpful Person:

- In addition to the team presentations, the Most Helpful Person award will be awarded to the most helpful person at the event, as voted upon by the hackathon participants.
- Before end of Saturday, participants will vote for the most helpful person. The person with the most votes will receive the Most Helpful Person award.
- Each participant will be given fixed number of points that they can distribute as they like to any other participant, except themselves. The participant with the most points will be awarded the prize.



DATASCIENCEGO

HACKATHON