Baker Hughes

DS Hackathon 2022

Visualization

March 2022

Challenge Description

As an energy technology company, at Baker Hughes (BH) we approach to a sustainable energy future by deploying the most efficient and least emissive technologies. To help industry advance on the path to net-zero and a sustainable energy future, one of the strategies to follow is identify, control and reduce emissions from operations.

In this challenge, you are given a synthetic dataset regarding Aeroderivative Gas Turbine operation, one of our most important technologies. You are asked to assume the role of a BH data scientist with the task to communicate valuable information about the emissions from gas turbine engines around the world.

Objectives

- The main objective is to produce a report, dashboard, or a set of visualizations with a great storytelling that allows you to clearly communicate your findings. For this task, you are free to subset as you choose the:
  - **Emissions theme**: worst cases, best cases, timeline report, etc.
  - **Level of analysis**: gas turbine, site, customer, etc.
  - **Type of charts**: bar, lines, scatter, pie, map, KPIs, gauge, treemap, etc.
  - **Functionality of visualizations**: interactive, responsive, etc.

- In addition, we are asking you to compute the estimated operating hours for a particular gas turbine engine and submit it in this challenge website.

Let's code and take energy forward! 🏠

Data Description

In this challenge we provide a synthetic dataset for the operation of gas turbine engines from different customers sites around the world.

As mentioned in Challenge Description, we are requesting the submission for scoring of the operating hours estimation (HOURS attribute) for engine 3 in the Ancient Wasp site from our Busy customer:

- HOURS: operating hours in h. This attribute is the accumulated operation time based on the measurements of the compressor speed. The gas turbine engine is considered to be running when two consecutive measurements of the compressor speed are greater than zero, then this sampling time is added to the operating hours count. The starting value of HOURS for all gas turbine engines is zero. Please assume that the gas turbine engines reach a non-zero speed instantly.

Files

- site_metadata.csv - geographical information of the sites, and the fuel used at them.
- engine_metadata.csv - information about the gas turbine customers, sites and their corresponding data files.
- dataset.zip - data collected from gas turbine engines as defined in engine_metadata.csv file.
- sample_submission.csv - a sample submission file in the correct format.

Data Description

Columns

Columns in site_metadata.csv:

- CUSTOMER_NAME: name of customer.
- PLANT_NAME: name of site.
- LATITUDE: in degrees.
- LONGITUDE: in degrees.
- ELEVATION: in meters.
- FUEL_LHV: lower heating value of the fuel in BTU/lb.

Columns in engine_metadata.csv:

- CUSTOMER_NAME: name of customer.
- PLANT_NAME: name of site.
- ENGINE_ID: engine name.
- FILE_ID: filename with data collected from gas turbine engines.

Columns in files data_#.csv included in dataset.zip:

- Date: datetime of measurement.
- CMP_SPEED: compressor speed in RPM.
- POWER: power output from the Low-Pressure Turbine (LPT) in kW.
- FUEL_FLOW: fuel flow into the combustor in kg/s.
- CO2: carbon dioxide estimated emissions in kg/s.

Columns in sample_submission.csv:

- ID: measurement id.
- HOURS: calculated operating hours, h.

A parameter that is usually calculated for gas turbine engines that might be useful in your analysis is:

- THRM_EFF: % of thermal efficiency.

$$\text{THRM\_EFF} = \frac{\text{POWER}}{\text{FUEL FLOW} * \text{FUEL LHV}} * 100$$

Evaluation

Data is a big world, it can be of several types, from different sources and have specific physical meanings; there is not always a set of defined tasks or steps to follow to retrieve valuable information from it. As professionals dedicated to Data Science, we must develop skills to optimally explore data and decide on what are the important leads to follow.

Creativity and good programming skills are two of the main qualities required for data scientists during data exploration, and in the selection of the analysis approach to solve a problem. It is also crucial to present and communicate your results to the people responsible of decision making. In this challenge we would like you to demonstrate these abilities.

As explained in the "Hackathon Rules and Guidelines" Evaluation Criteria, the evaluation is divided into two equal parts: team presentation, and technical approach.

For the **Team presentation** we will evaluate considering the "Hackathon Rules and Guidelines" Evaluation Criteria. Keep in mind that the length of the presentation is limited to 30 minutes.

For the **Technical approach**, we expect you to explore the given data and tell a compelling story through the visualizations of your preference, clearly stating the problem or questions that your visualizations will help to answer. We will consider:

- **Clarity** of the analyzed data and its relevance to the proposed problem/question.
- **Accuracy** of the graphical representation used to convey your message.
- Efficiency of **visual effects** (use of appropriate shapes, colors and sizes to represent the analyzed data).
- **General evaluation** attributes: orthography, originality, authenticity, etc.
- **Computational and software resources** selected:
  - **Kaggle score leaderboard** (*next slide*)
  - **Code submission** (with HOURS estimation and visualizations development)

  The highest possible score can only be achieved using Python-Kaggle resources. If you use other languages, you will be given 80% of the points.

For this challenge, along with the presentation you must submit the developed product (report, dashboard, or a set of visualizations) in pdf format.

Baker Hughes

Evaluation

Kaggle score leaderboard

Submit your computation of the operating hours estimation for engine 3 in the Ancient Wasp site from our Busy customer, as described in Data.

- Metric

  The evaluation metric for the estimation is the Root Mean Square Error (RMSE). As your submission approaches the solution, the Kaggle score is near to zero.

- Submission Format

  Your csv file should contain two columns: ID and HOURS. Where ID are non-negative integers uniquely identifying the measurements taken from the gas turbine engine, and `HOURS` is your estimation for the operating hours for Busy's engine 3 at Ancient Wasp, in h. A submission example is available at Data. The file must contain a header with the following format:

```
ID,HOURS
0,0
1,0
2,1
3,2
4,2
...
8760,5498
```

Evaluation

Challenge evaluation score table:

Evaluation parts	Evaluation attributes		Point score
Team presentation (45 points)	Time management		5
	Speech quality		5
	Clarity of findings and results		15
	Quality of presentation material		5
	Ability to summarize		15
Technical approach (55 points)	Clarity		10
	Accuracy		10
	Visual effect		8
	General Evaluation		7
	Computational and software resources	Kaggle score leaderboard	3
		Code submission	17
		Total	**100**

Consider:

- You will be given zero points in the technical approach evaluation if your team did not submit your Kaggle HOURS estimation and the code by **March 12, 2022 11:00 pm (CDMX time)**.
- You will be given zero points in the team presentation if your team did not submit your presentation material by **March 13, 2022 08:00 am (CDMX time)**.

FAQs

Can we use any programming language?
Yes, but Python is the preferred language. Using other language will give you at much the 80% of the Code submission Point score. Don't forget that you must send your code for evaluation on time, no matter what language you used.

How are we going to send the code?
GitHub is going to be the official platform for all deliverables. *Consult the provided material to interact with Kaggle from GitHub*. You must create a GitHub repository as a team and push the commit of your final code by the same deadline as the Kaggle submission: **March 12, 2022 11:00 pm (CDMX time)**. Use the form to share the GitHub repository link for the judges to evaluate all deliverables.

How are we going to send the presentation material, and the developed product (report, dashboard, or a set of visualizations)?
GitHub is going to be the official platform for all deliverables. In the same GitHub repository of your code, you must upload the presentation, and the product (in pdf format) with deadline: **March 13, 2022 08:00 am (CDMX time).** Use the form to share the GitHub repository link for the judges to evaluate all deliverables.

What should the presentation material include?
Remember that you are assuming the role of a BH data scientist, working in a Data Science Project with the task to deliver a Data Product about the emissions from gas turbine engines around the world. Here are some references: How to Present Your Data Science Project, Creating a Presentation for a Data Science Project, 4 Tips to Boost Your Data Science Project Presentation.

How to win the challenge?
The team with the highest overall score wins the TPS Visualization Challenge! (page 7 *Challenge evaluation score table*)

For more questions, wait for the **Open hours with mentors** or reach us via the authorized **Slack** communication channels.

Baker Hughes

References

Most popular Python libraries for visualization:

- Matplotlib
- Seaborn
- Plotly
- Jupyter Widgets
- Folium

Improve your skills in visualization and data storytelling:

- Data Visualization Tutorial For Beginners (with python) by Simplilearn
- Storytelling with data chart guide
- Create An Infographic Using Matplotlib
- Report examples
- Infographic example
- Dashboard examples
- Data visualization MOOCs:
  - Data Visualization with Python by IBM
  - Fundamentals of Data Visualization by University of Colorado
- Data visualization talks-books:
  - Storytelling with Data by Cole Nussbaumer Knaflic talk book
  - Designing Data Visualizations by Noah Iliinsky talk book

About Baker Hughes:

- We are Baker Hughes, an energy technology company
- Energy transition
- The path to net-zero and a sustainable energy future
  - Episode 1
  - Episode 2

About kaggle:

- Kaggle's competition documentation
- Instructions on submitting predictions