Applied Tech. Project 67 - Car Prices Prediction - Data Exploration

October 29, 2020

0.0.1 Instructions

Goal of the Project This project is designed for you to practice and solve the activities that are based on the concepts covered in the following lessons:

1. Multiple linear regression - Introduction
2. Multicollinearity
3. Variance Inflation Factor

Getting Started:

1. Click on this link to open the Colab file for this project.

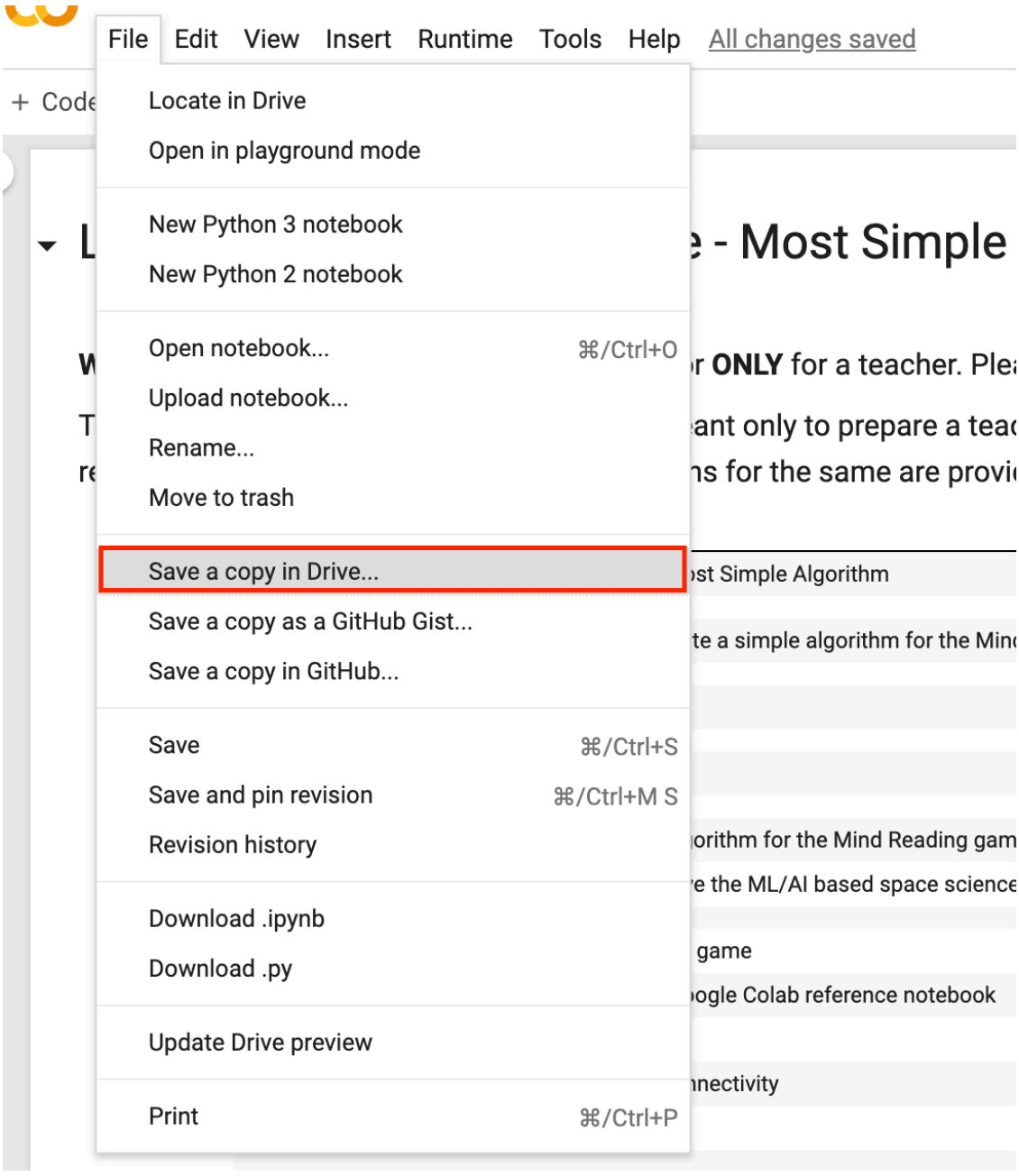
https://colab.research.google.com/drive/1niw1N9m_B8RuLfVwrh-WmKyA7hBU_qDS?usp=sharing

2. Create a duplicate copy of the Colab file as described below.

- Click on the **File menu**. A new drop-down list will appear.

!

- Click on the **Save a copy in Drive** option. A duplicate copy will get created. It will open up in the new tab on your web browser.



3. After creating the duplicate copy of the notebook, please rename it in the **YYYY-MM-DD_StudentName_Project67** format.
4. Now, write your code in the prescribed code cells.

0.0.2 Problem Statement

Implement multiple linear regression to create a predictive model capable of predicting the yearly amount spent by the customers in shopping from an Ecommerce website. Find out if there exists multicollinearity in the dataset using Variance Inflation Factor.

0.0.3 List of Activities

Activity 1: Analysing the Dataset

Activity 2: Exploratory Data Analysis

Activity 3: Train-Test Split

Activity 4: Model Training using `statsmodels.api`

Activity 5: Calculate VIF using `variance_inflation_factor`

Activity 1: Analysing the Dataset

- Create a Pandas DataFrame for **ecommerce-customers** dataset using the below link. This dataset consists of following columns:

Columns	Description
Email	Email ID of the customer
Address	Address of the customer
Avatar	Color Avatar
Avg. Session Length	Average session of in-store style advice sessions
Time on App	Average time spent on App in minutes
Time on Website	Average time spent on Website in minutes
Length of Membership	How many years the customer has been a member
Yearly Amount Spent	Amount spent on items yearly

Dataset Link: https://student-datasets-bucket.s3.ap-south-1.amazonaws.com/whitehat-ds-datasets/ecommerce-customers.csv

- Print the first five rows of the dataset. Check for null values and treat them accordingly.
- Also, drop the columns `Email`, `Address` and `Avatar` as they are not required for further analysis.

[ ]:
```
# Import modules

# Load the dataset

# Print first five rows using head() function
```

[ ]:
```
# Check if there are any null values. If any column has null values, treat them␣
↪accordingly
```

[ ]:
```
# Drop unnecessary columns
```

Activity 2: Exploratory Data Analysis Create the scatter plots between each independent variables and the target variable. Determine which independent variable(s) shows linear relationship with the target variable `Yearly Amount Spent`.

[ ]:
```
# Create scatter plot with 'Avg. Session Length' on X-axis and 'Yearly Amount␣
↪Spent' on Y-axis
```

[ ]:
```
# Create scatter plot with 'Time on App' on X-axis and 'Yearly Amount Spent' on␣
↪Y-axis
```

[ ]:
```
# Create scatter plot with 'Time on Website' on X-axis and 'Yearly Amount␣
↪Spent' on Y-axis
```

[ ]:
```
# Create scatter plot with 'Length of Membership' on X-axis and 'Yearly Amount␣
↪Spent' on Y-axis
```

Q: Based on the scatter plots, which independent variable seems to have the best linear relationship with the target variable?

A:

Activity 3: Train-Test Split We need to predict the value of `Yearly Amount Spent` variable, using other variables. Thus, `Yearly Amount Spent` is the target or dependent variable and other columns except `Yearly Amount Spent` are the features or the independent variables.

Split the dataset into training set and test set such that the training set contains 70% of the instances and the remaining instances will become the test set.

[ ]:
```
# Split the DataFrame into the training and test sets.
```

Activity 4: Model Training using `statsmodels.api` Perform the following tasks: - Implement multiple linear regression using `statsmodels.api` module and find the values of all the regression coefficients using this module. -Print the statistical summary of the regression model.

[ ]:
```
# Build a linear regression model using the 'statsmodels.api' module.


  # Add a constant to feature variables


  # Fit the regression line using 'OLS'


  # Print the parameters, i.e. the intercept and the slope of the regression␣
↪line fitted
```

```
[ ]: # Print statistical summary of the model
```

Q: What is the R^2 (R-squared) value for this model?

A:

Activity 5: Calculate VIF using `variance_inflation_factor` Calculate the VIF values for each independent variables using the `variance_inflation_factor` function of the `statsmodels.stats.outliers_influence` module.

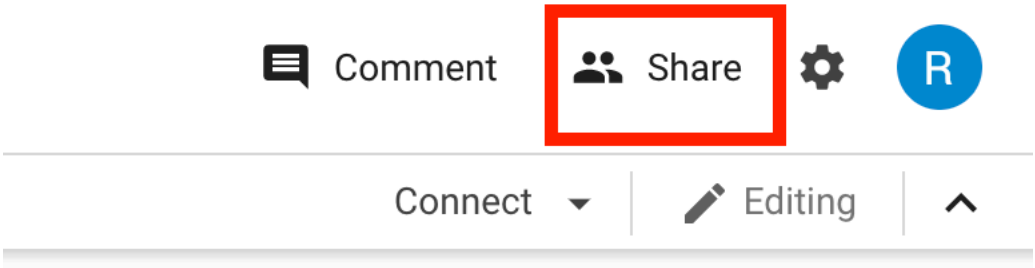
```
[ ]: # Calculate the VIF values for each independent variable using the␣
     ↪'variance_inflation_factor' function.


     # Create a dataframe that will contain the names of all the feature variables␣
     ↪and their respective VIFs
```

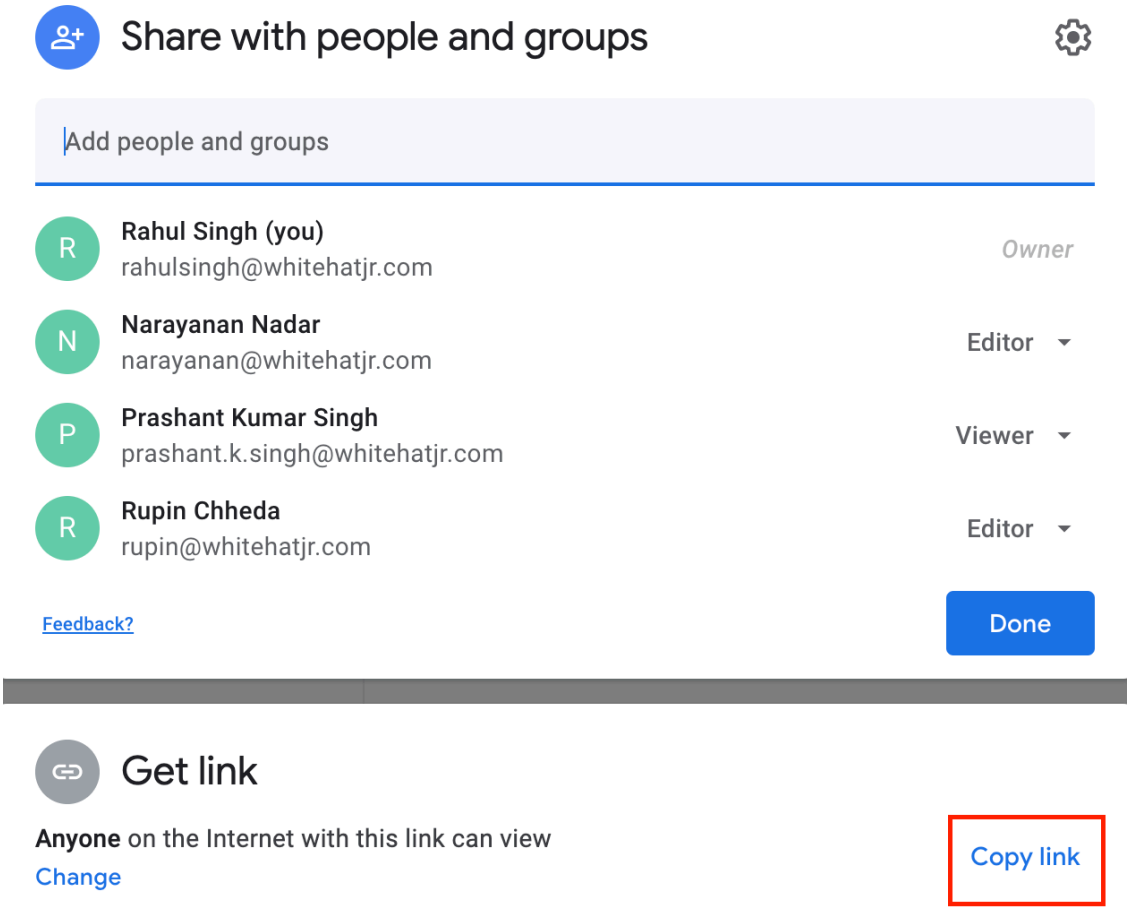
---

1 Submitting the Project:

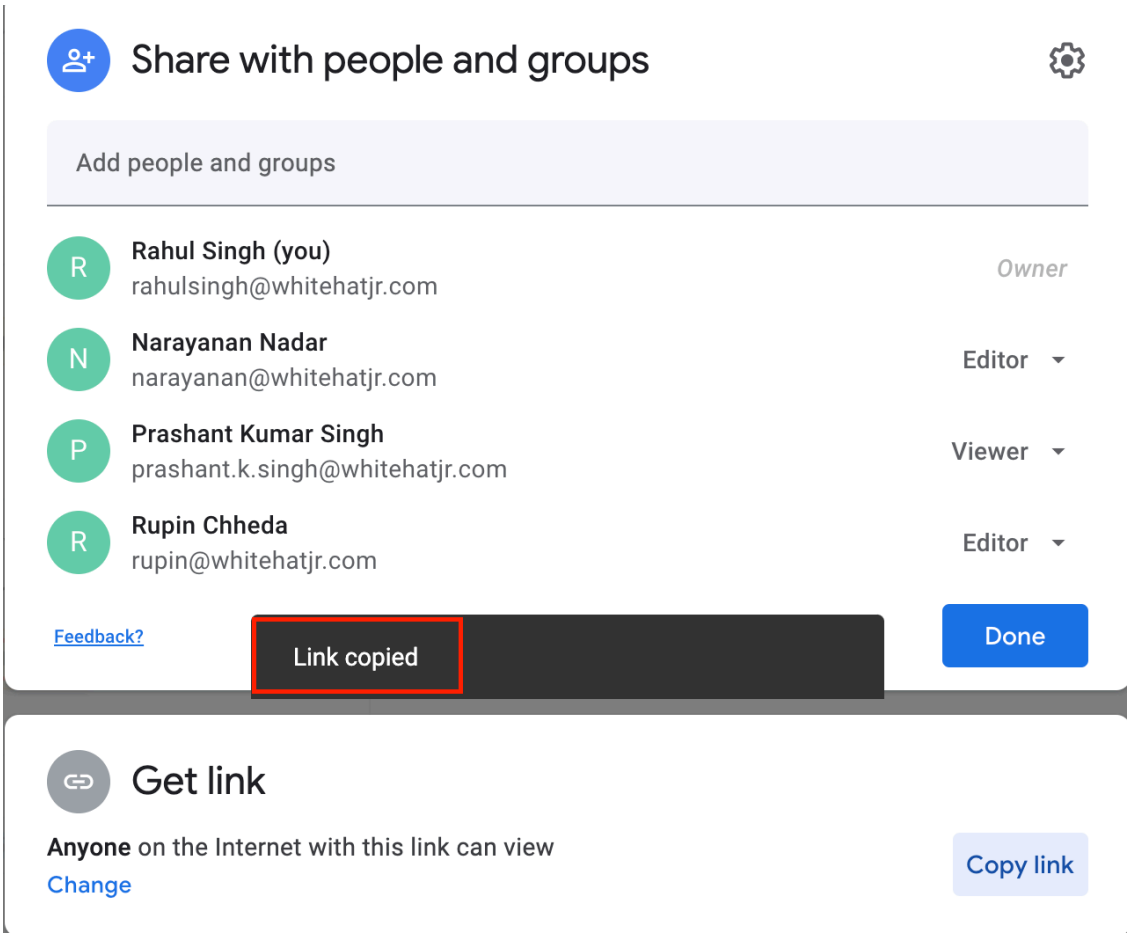
1. After finishing the project, click on the **Share** button on the top right corner of the notebook. A new dialog box will appear.



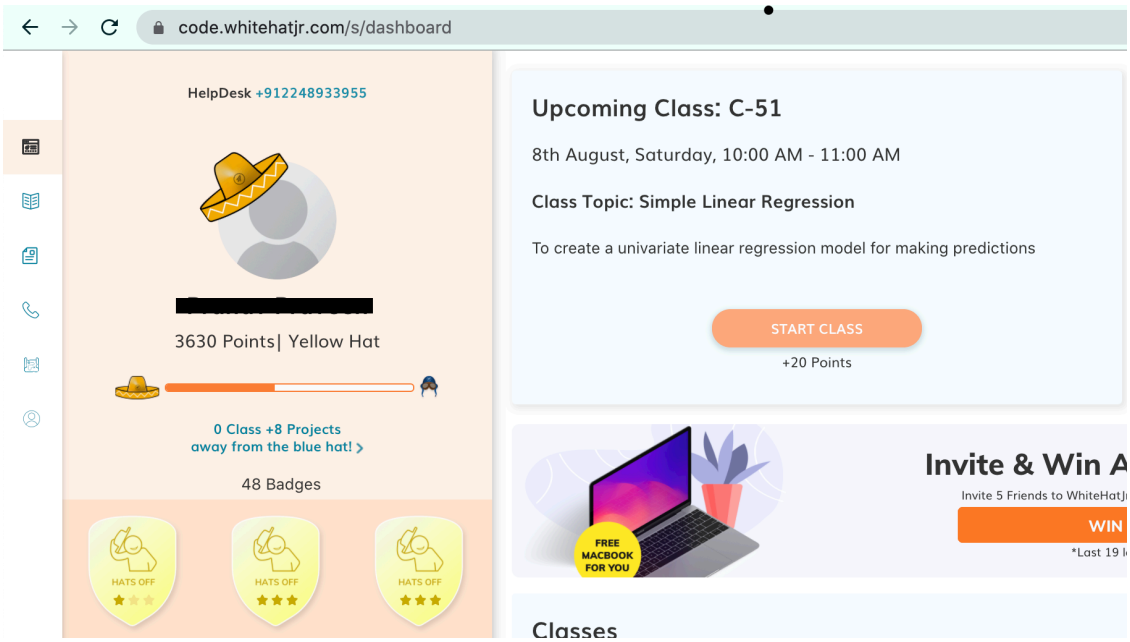
2. In the dialog box, make sure that '**Anyone on the Internet with this link can view**' option is selected and then click on the **Copy link** button.

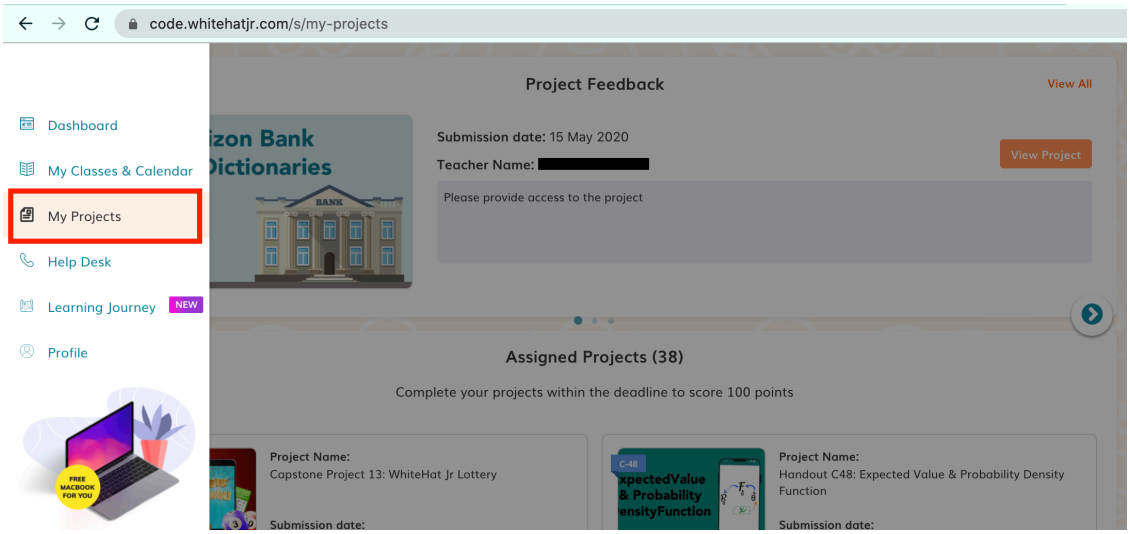


3. The link of the duplicate copy (named as **YYYY-MM-DD_StudentName_Project67**) of the notebook will get copied

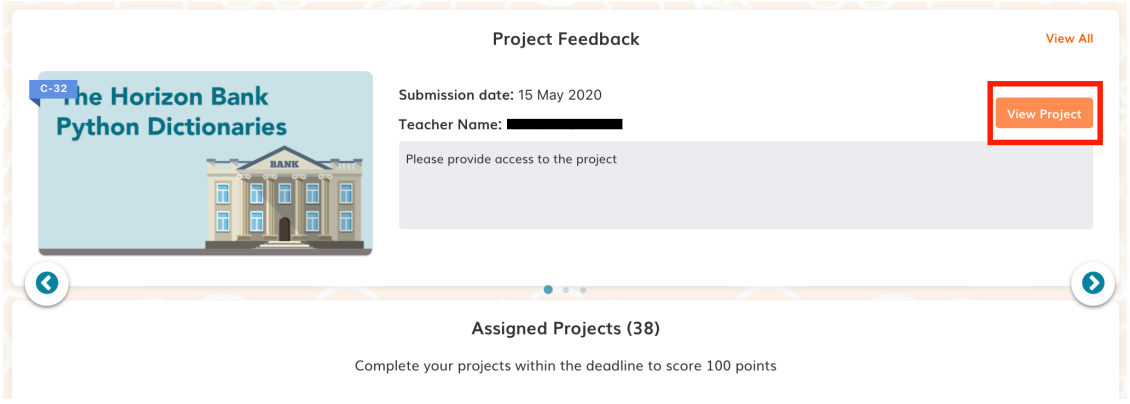


4. Go to your dashboard and click on the **My Projects** option.

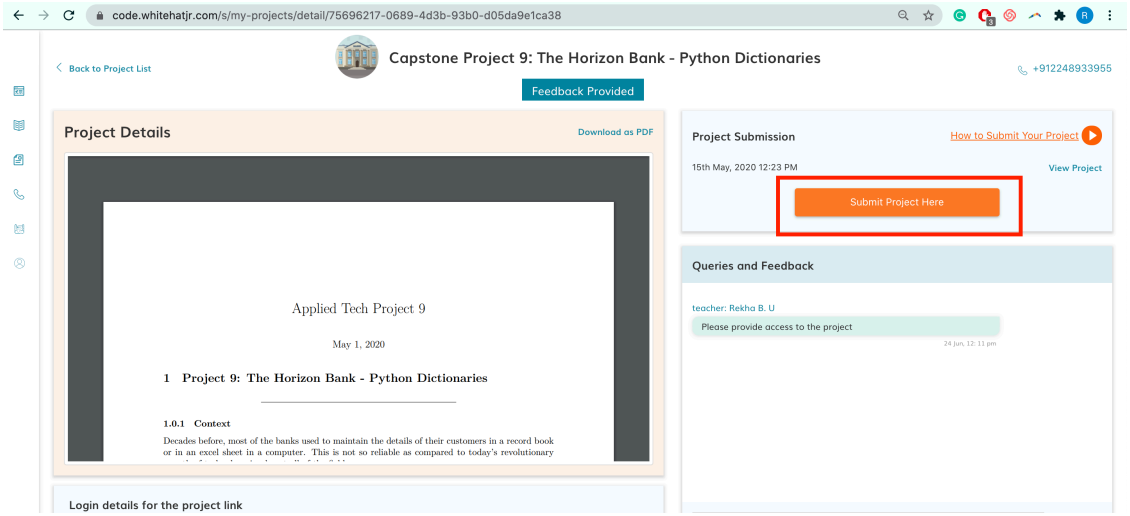




5. Click on the **View Project** button for the project you want to submit.



6. Click on the **Submit Project Here** button.



7. Paste the link to the project file named as **YYYY-MM-DD_StudentName_Project67**

in the URL box and then click on the **Submit** button.