8/22/2019

Simran Chhabra

for Edwisor.com

Employee Absenteeism

Contents

1. **Introduction**
   1. Problem Statement
   2. Data
2. **Methodology**
   1. Pre-Processing
      1. Outlier Analysis
      2. Missing values
      3. Feature Selection
   2. Modelling
      1. Dependency Check
      2. Linear Regression
      3. Random Forest Regressor
      4. K Neighbours Regressor
3. **Conclusion**
   1. Model Evaluation
   2. Model Selection
   3. Answers

**Appendix A – Python Code**

**Appendix B – R Code**

**References**

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. The company has shared it dataset and requested to have an answer on the following areas:

1. What changes company should bring to reduce the number of absenteeism?

2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

**1.2 Data**

Dataset Characteristics: Timeseries Multivariant

**Attribute Information:**

1. Individual identification (ID)

2. Reason for absence (ICD)

3. Month of absence

4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

5. Seasons (summer (1), autumn (2), winter (3), spring (4))

6. Transportation expense

7. Distance from Residence to Work (kilometers)

8. Service time

9. Age

10. Work load Average/day

11. Hit target

12. Disciplinary failure (yes=1; no=0)

13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

14. Son (number of children)

15. Social drinker (yes=1; no=0)

16. Social smoker (yes=1; no=0)

17. Pet (number of pet)

18. Weight

19. Height

20. Body mass index

21. Absenteeism time in hours (target)

**Chapter 2**

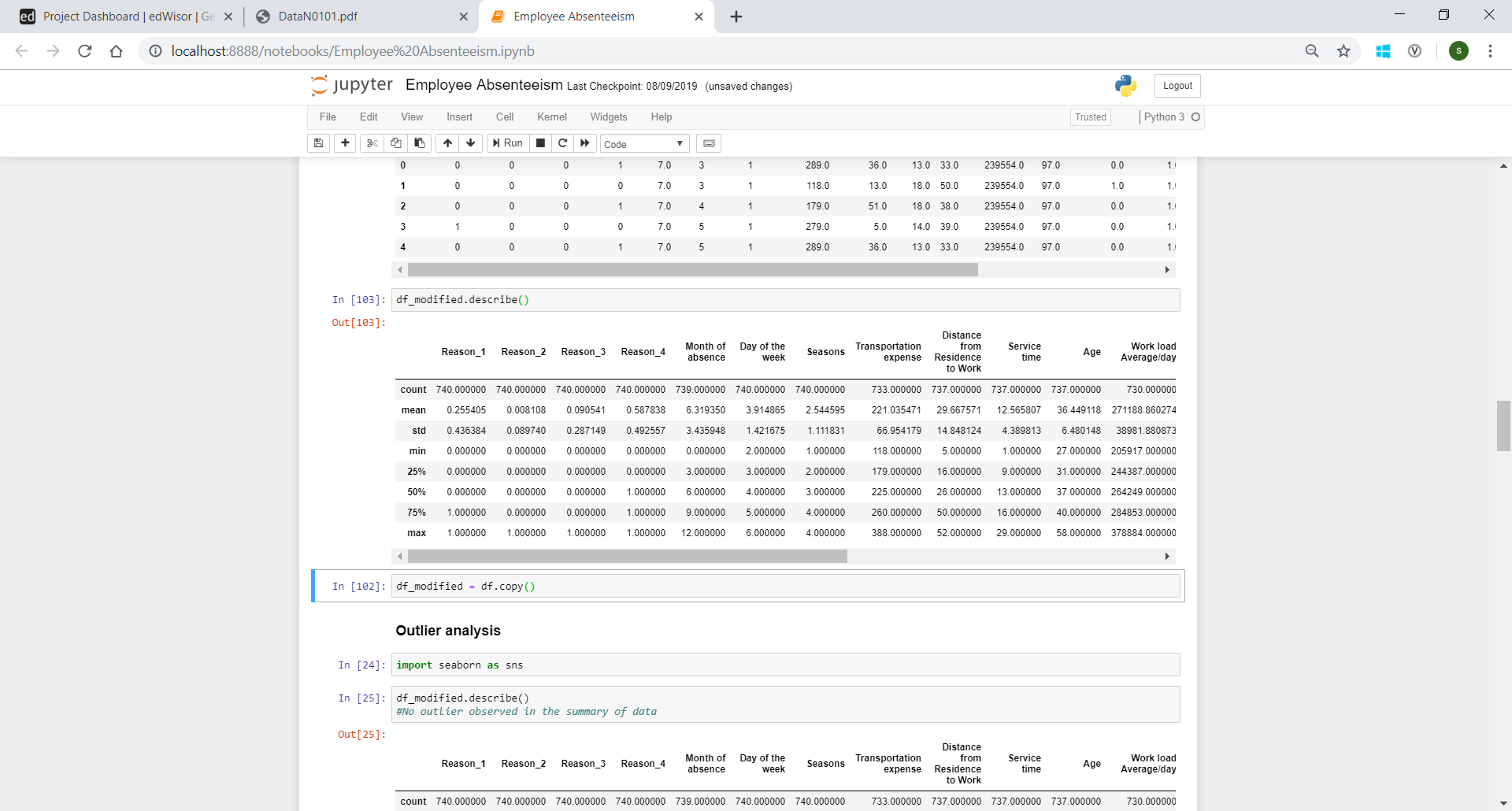
**Methodology**

* 1. **Pre-Processing**

Any predictive modelling requires that we look at the data before we start modelling. However, in Data Mining terms *looking at data* means so much more than looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis.

* To start with, we first check the data at high level. It was observed that we have been given columns ‘Weight’, ‘Height’ and ‘Body mass index’. Since BMI is calculated using weight and height itself, we do not need these two columns. Also, the unique identifier field ‘ID’ would in no way help us in predicting absenteeism thus that field could also be waved off. So ‘ID’, ‘Weight’ and ‘Height’ fields were dropped in the initial step.
* Next, we start check each attribute one by one. The second attribute ‘Reason for absence’ is a categorical attribute with 28 unique values. To analyse the reasons given by employees in depth, we need to separate these categories into columns. Thus, 28 dummy variables were created. These variables were then reduced to 4 variables as they were related as below:
  + Categories 1 to 14 were related to some infection or disease.
  + Categories 15 to 17 were related to pregnancy.
  + Categories 18 to 21 were related to some injury and external causes
  + Categories 22 to 28 were related to visiting a doctor/consultation.
    1. **Outlier Analysis**

On checking the data’s summary (refer figure 1), there were no such outliers observed. Maximum and minimum values of each variable seemed to be in range. We can also graph the data on boxplots and check for any outliers (refer figure 2). If found, they could be corrected by first finding maximum and minimum values for that variable then dropping any values not lying in the range of minimum and maximum.



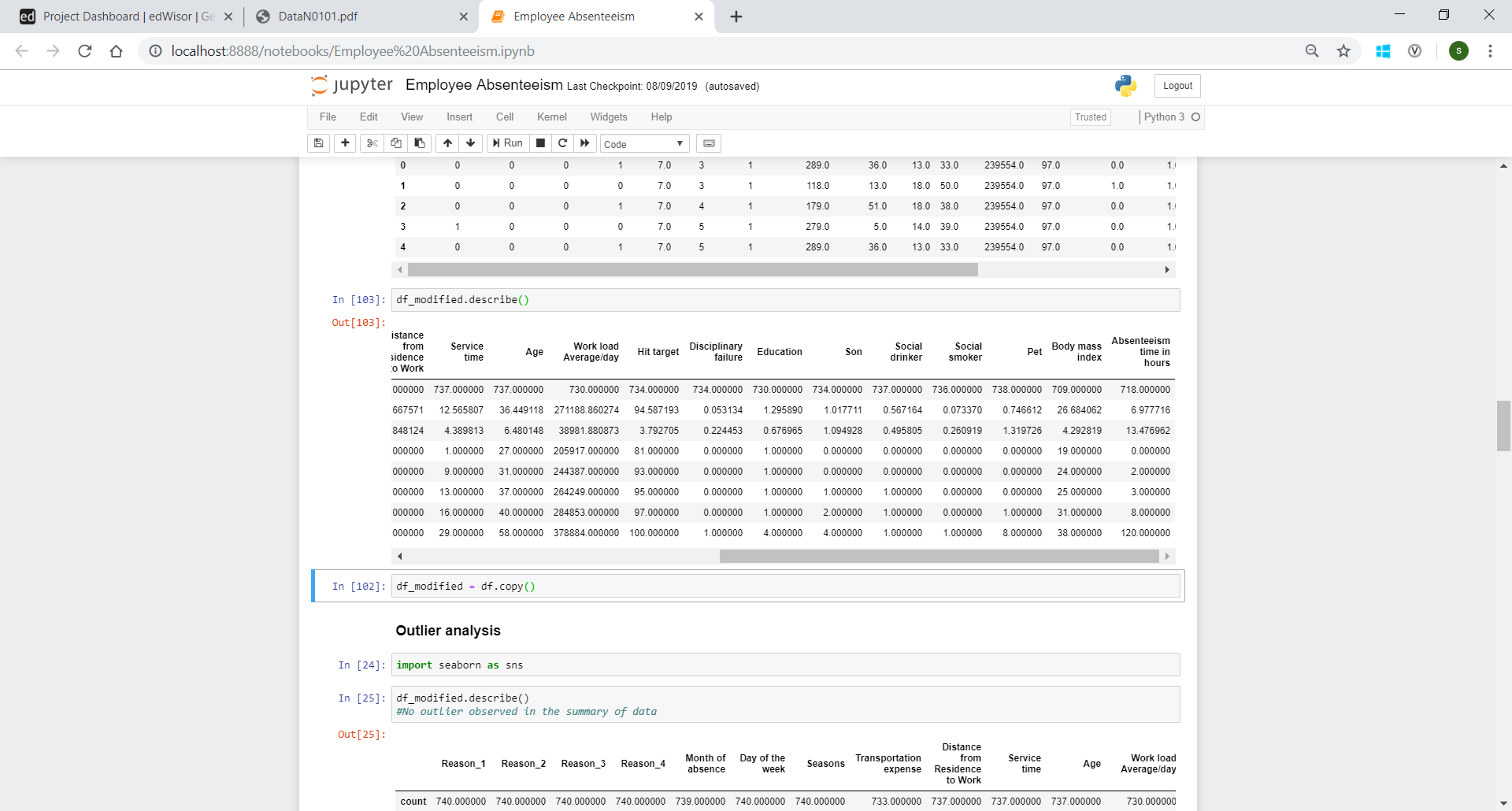


Figure 1

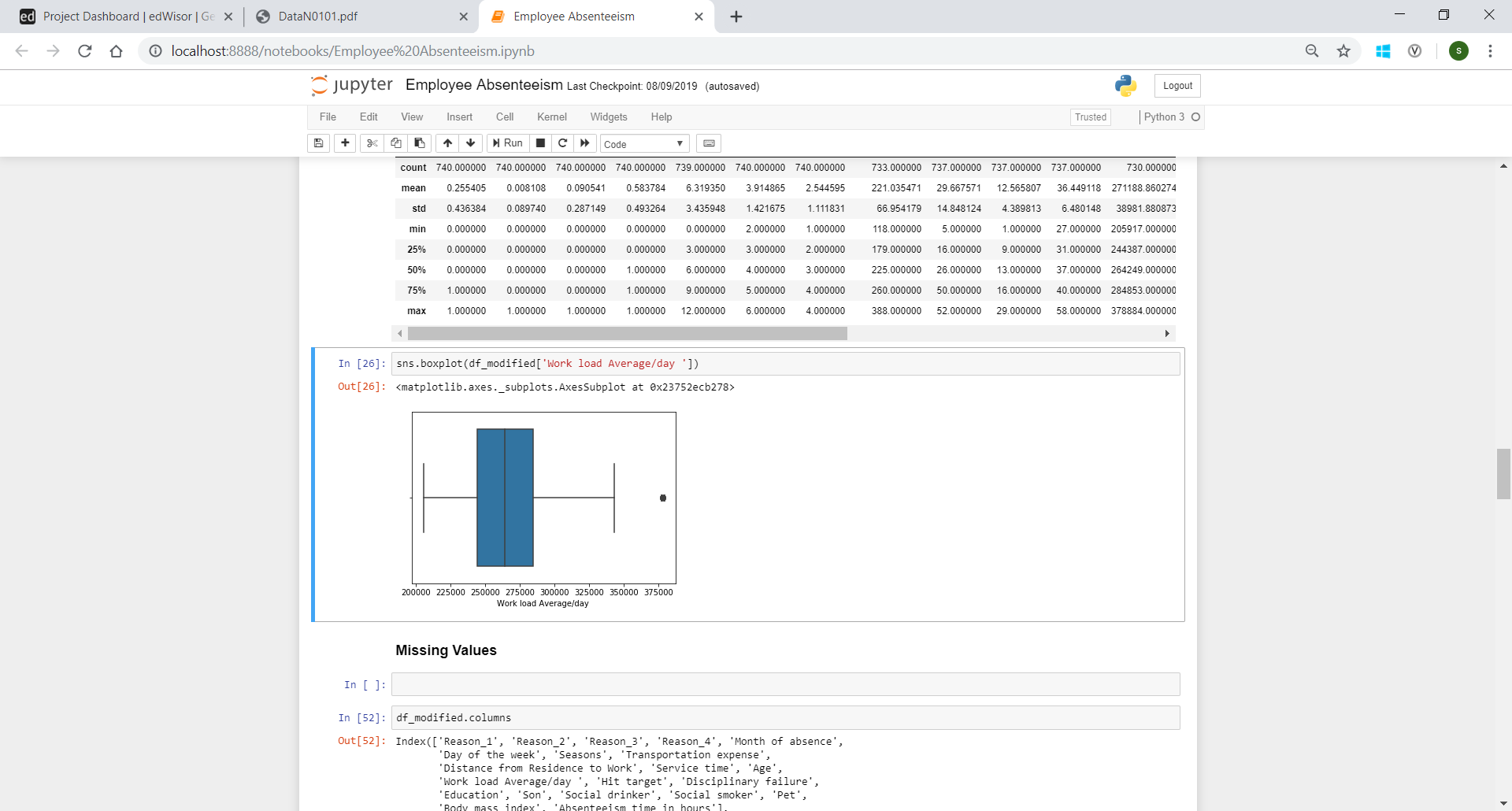


Figure 2

**2.1.2 Missing values**

Ideally if a variable has less than 30% missing values then they are filled using appropriate metrics like mean, median, mode or K-nearest neighbours. This will preserve the data and there will be no loss of information. Below are the percentage of missing values in each variable:

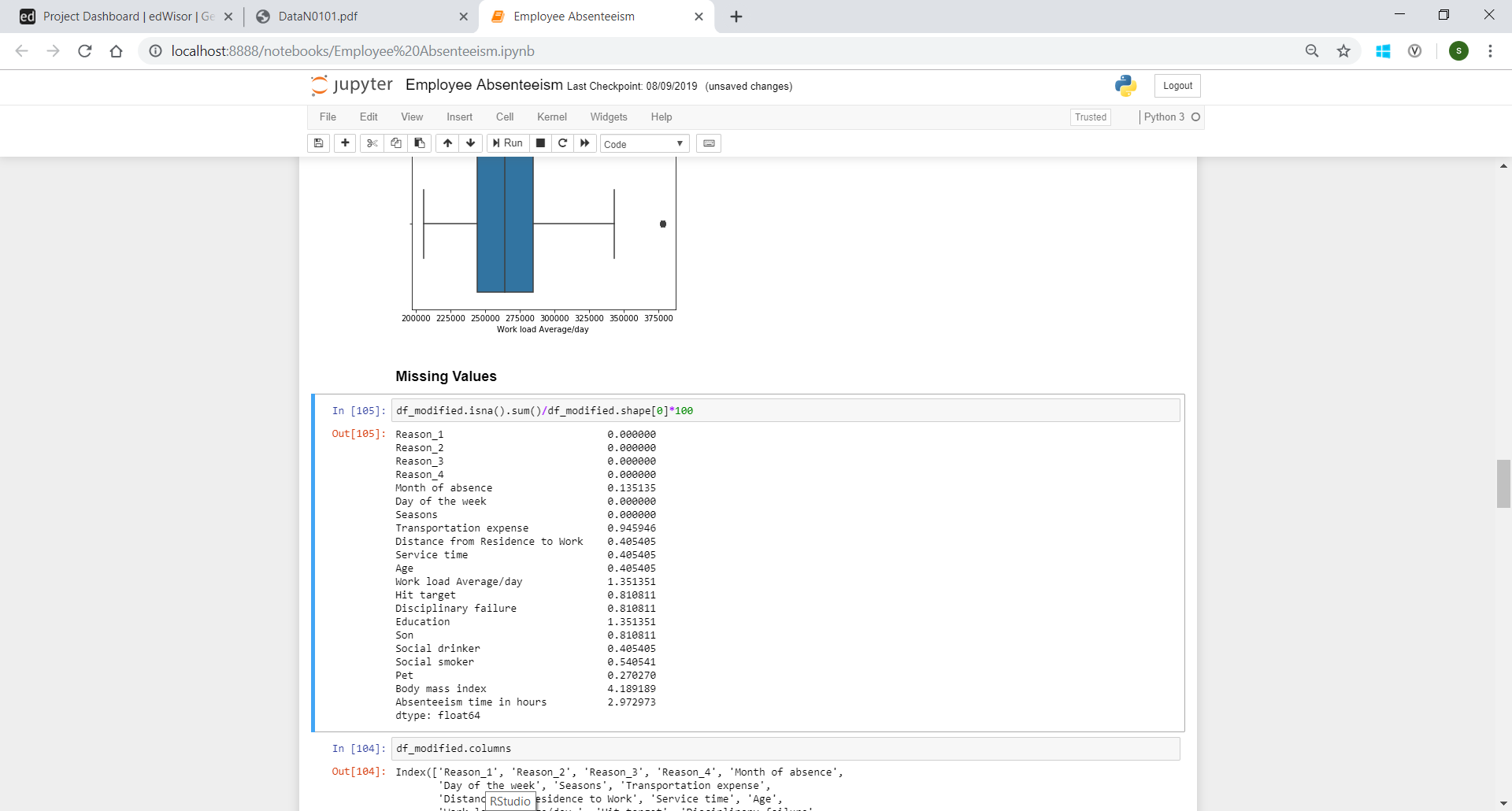


Figure 3

Since the missing values are less than 5% in each variable, they could be easily filled using mean/median (in case of numeric variable) and mode(in case of categorical variable). Other method could be KNN but since it is a time-consuming process and here not many values were blank, mean and mode were suitable to use.

* + 1. **Feature Selection**

Next, we focus on the memory space used by our data. In case of big data with huge number of columns, space and time become important factors. Thus, we try to reduce the size of our data by making sure that no information is lost. To do this, we check correlation between variables. If two variables are highly correlated, one of them could be discarded as they are eventually giving us similar information. Heatmap was created with to check correlation (refer Figure 4).

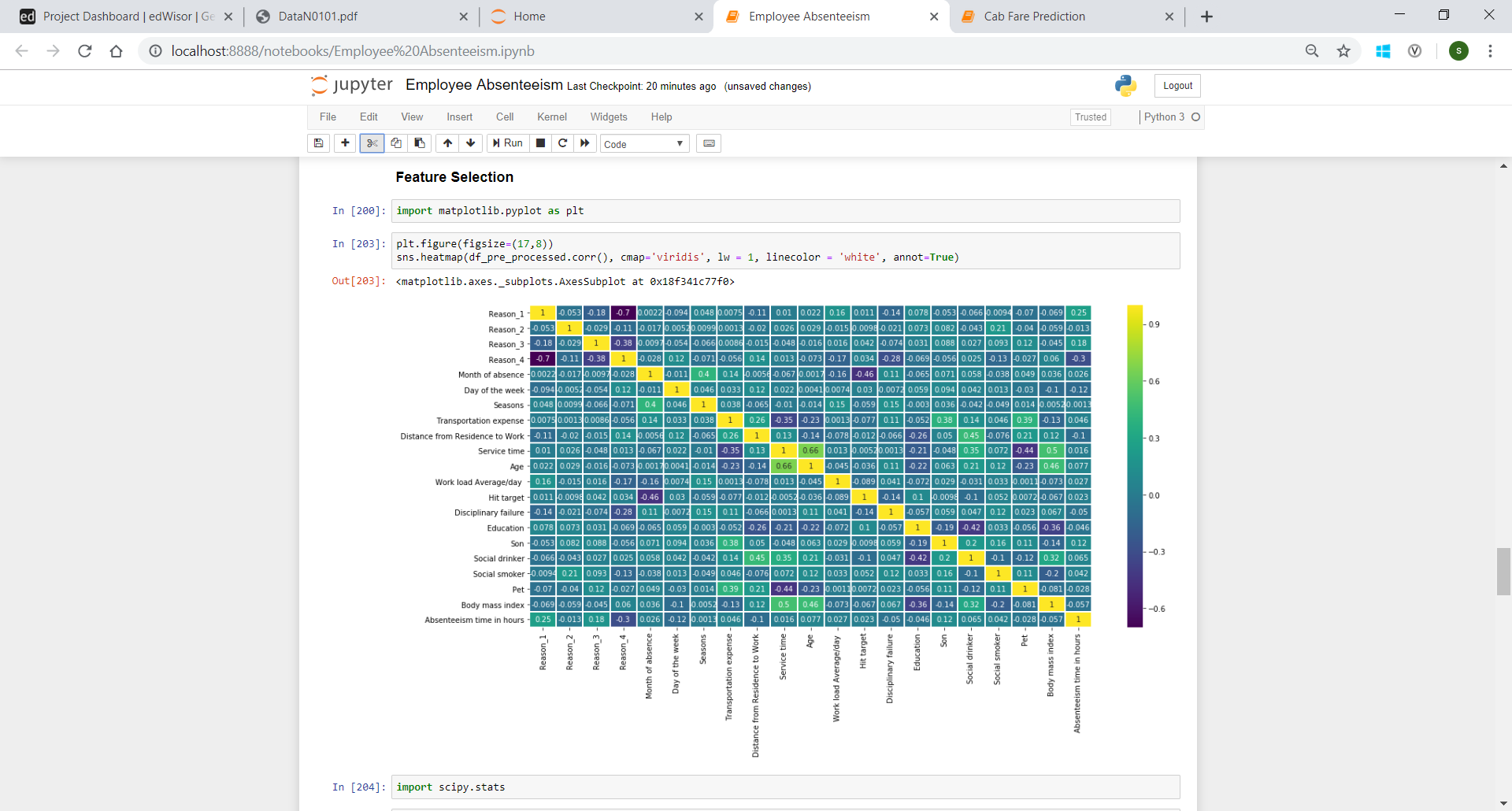


Figure 4

* 1. **Modelling**

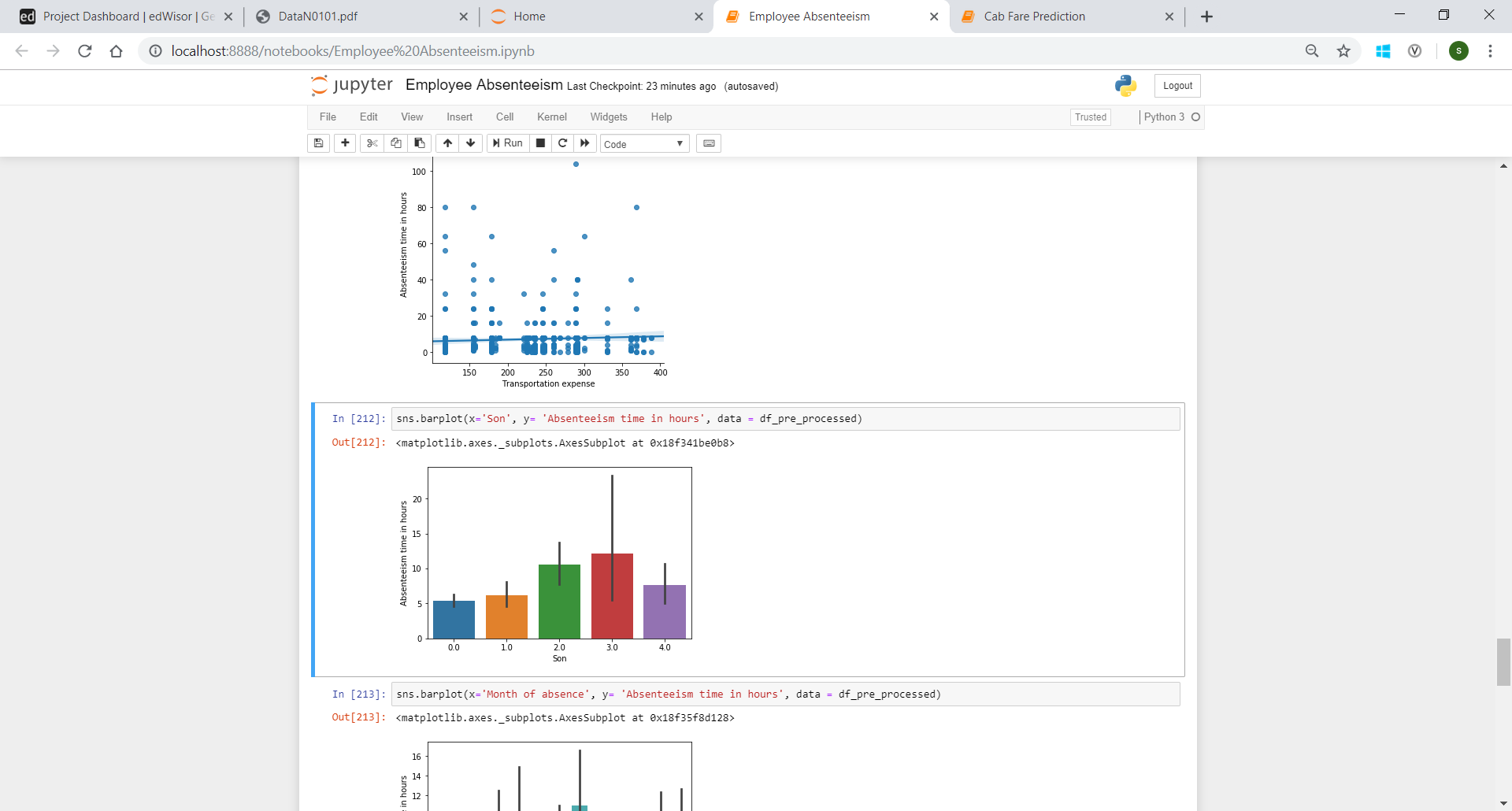
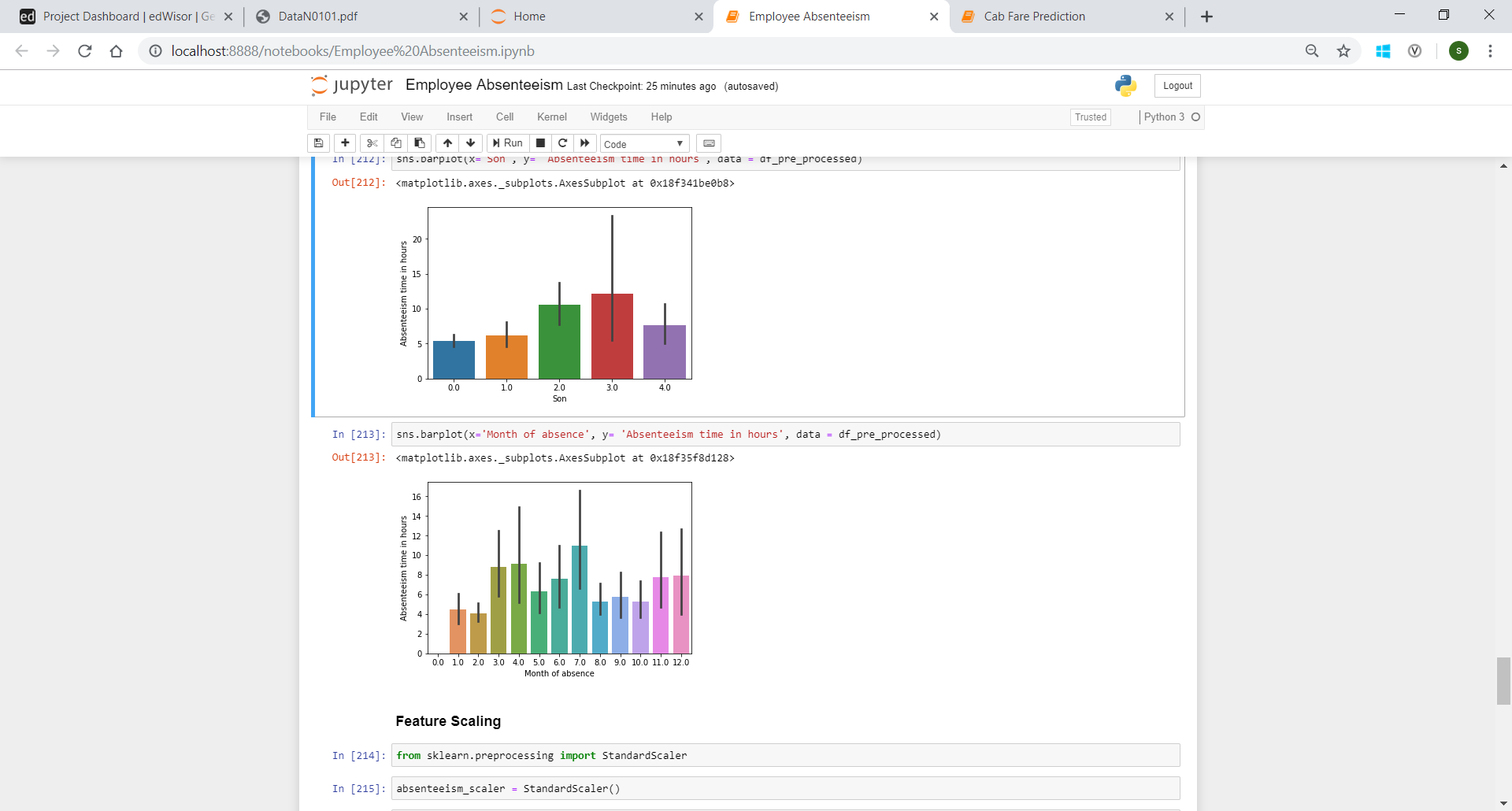
Heatmap (Figure 4) suggests that Absenteeism has higher dependency on fields Reason\_1, Reason\_3, Transportation expense, Month of absence, Age and Son than any other field.

Expected relations:

An employee is found to be on higher number of leaves when:

1. He has some infection or disease or some injury due to external causes.
2. His expense of transportation from work to home and back is high.
3. Month is one of April or July.
4. His Age is increasing.
5. He has many children.
   * 1. **Dependency check**

Visualize and check dependency of target variable on independent variables.

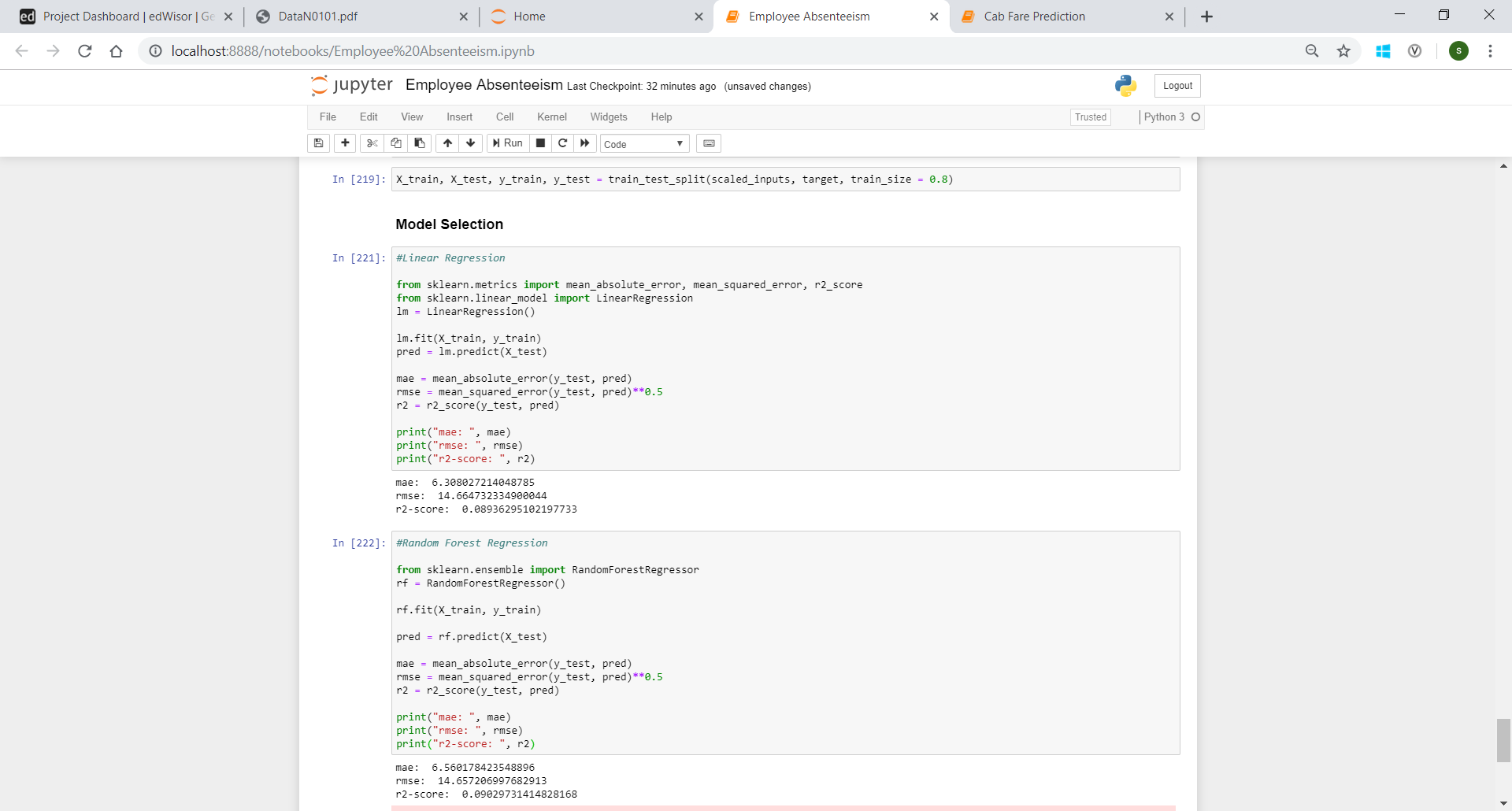
 

The data is first split into train and test data to predict. We move further to apply regression models on the train data to predict test values. The best model is picked up and then the values are predicted for the provided test data and error metrics are checked of each model.

* + 1. **Linear Regression**

We first check for assumptions needed for Linear regression to be performed that are: linear relationship, normal distribution, no multicollinearity and no auto-correlation. Data is scaled if not done already.

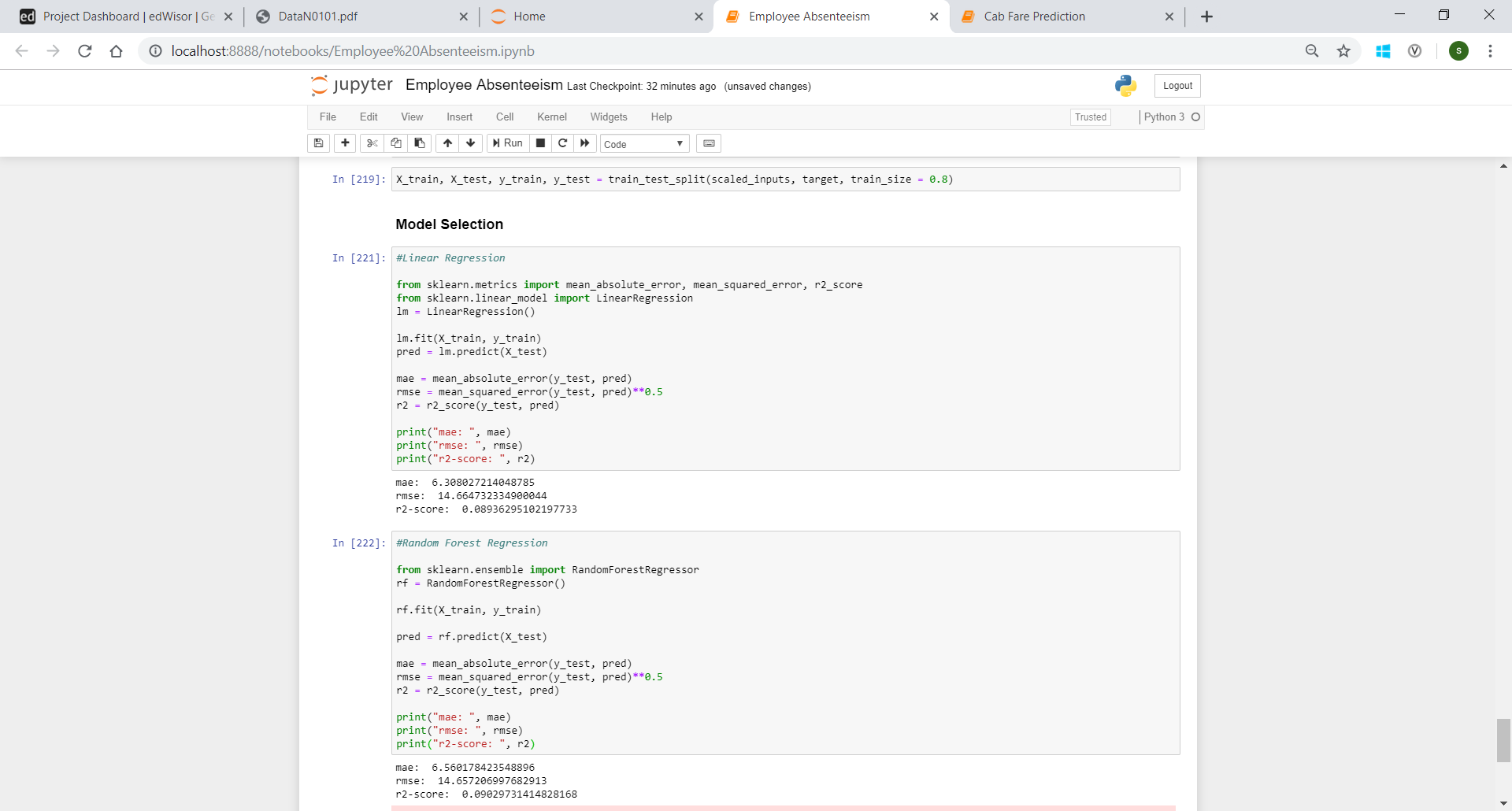
We get following error metrics on performing linear regression:



* + 1. **Random Forest Regression**

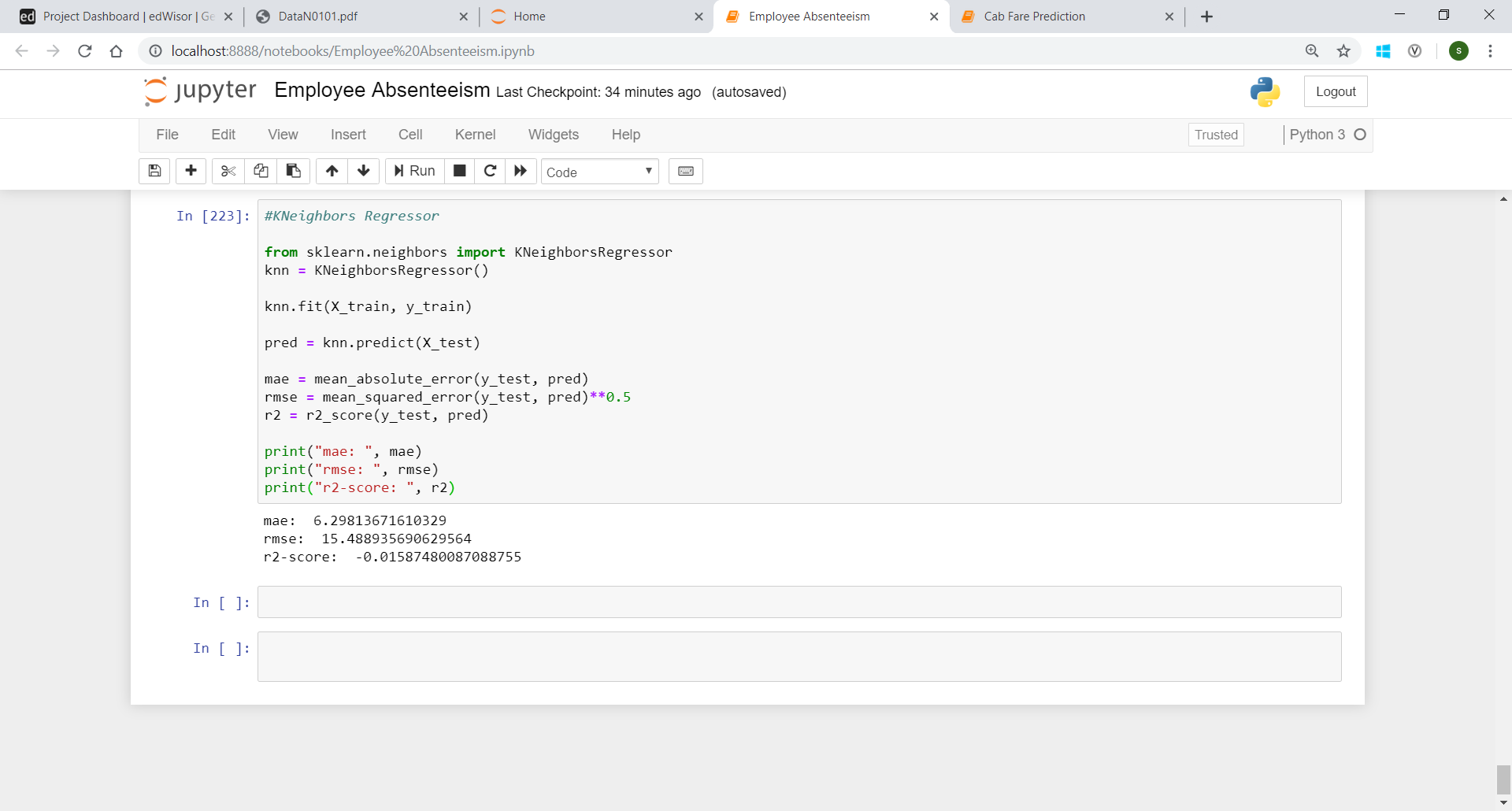
Now we try and use a different regression model to predict our target variable which Decision Tree regression.

We get following error metrics on performing it:



* + 1. **K Neighbours Regression**

Next, test values are predicted using K Neighbours regression and the following results are observed:



**Chapter 3**

**Conclusion**

* 1. **Model Evaluation**

Now that we have a few models for predicting the target variable, we need to decide which one to choose. We have used error metrics like mean absolute error (MAE), root mean squared error (RMSE) and R-squared (r2-score) to compare the performance of different models.

MAE and RMSE give us the error in the prediction values as compared to the actual values and r2-score tells us what proportion of variance in the target variable was successfully explained by the independent variables.

**3.2 Model Selection**

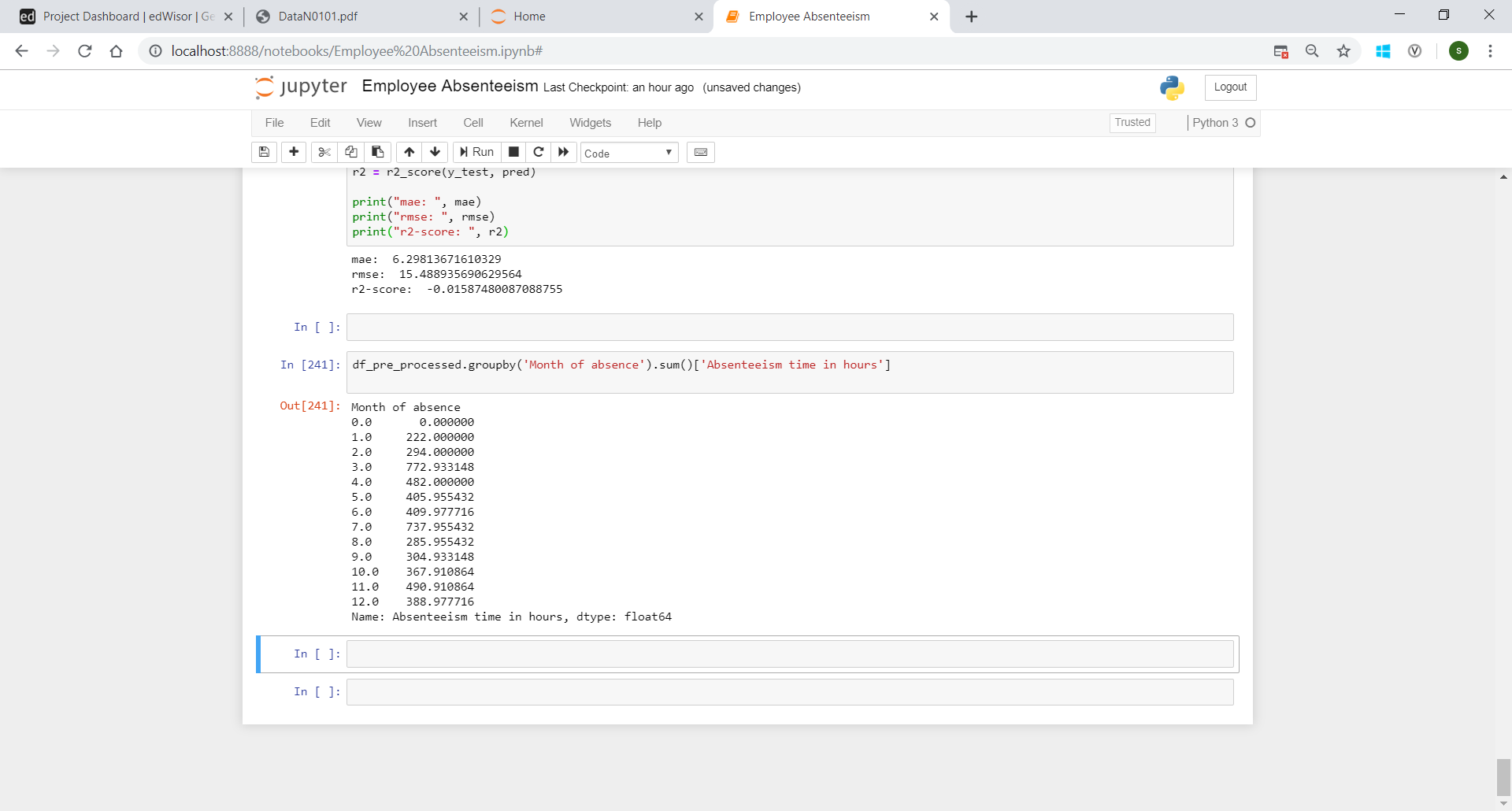
Less the value in error metrics and more the value in R-squared parameter, better is the model. Comparing these values in the three models, Linear Regression gives the least MAE and RMSE score and highest r2-score thus it would be the best suited model for our prediction.

* 1. **Answers**

To reduce the number of absenteeism, company can:

1. If the company is not already providing cab facilities, it should provide it at a minimal cost to all its employees so that people who live far away do not have to pay high fares in cab.
2. Provide facility of working from home in extremely hot, humid and cold weathers.
3. Provide mandatory paternal and maternal leaves to the employee. Employees with children less than age of 2 should also be allowed to leave early.

Here is the monthly distribution of Absenteeism time in hours:



**Appendix A**

Please find attached python notebook “Python\_Employee\_Absenteeism.ipynb” file.

**Appendix B**

Please find attached “R\_Employee\_absenteeism.R” file.

**References**

<https://learning.edwisor.com/>