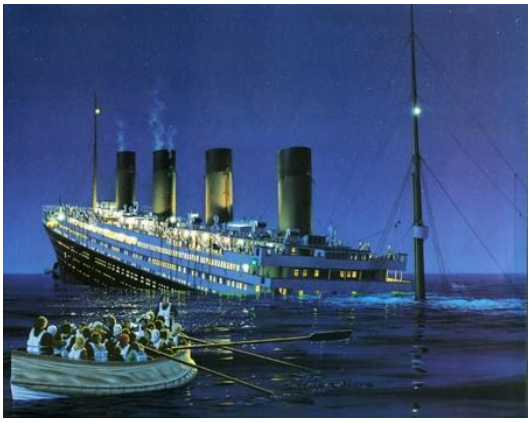
Data 650

Assignment 2

Instructions and Grading Rubric



For this assignment, you will use the Apache Spark to build and evaluate the performance of the logistic regression model for Titanic data and for the Low Birth Weight data. In addition to hands on parts, this assignment includes research questions.

Table of Contents

Getting Started

Contact your faculty immediately if you have less than 10 capacity unit hours remaining on your IBM cloud account for this month.

You may not work on this assignment until Apache Spark tutorial assignment is submitted for grading.

Do not delete the project that was setup for Spark Tutorial assignment. You may reuse that project for assignment 2.

Complete the readings in week 5-6 course content **before** working on assignment.

Download the following files from Assignment 2 folder in addition to this file. Save the files on your PC.

Python Notebook for Part 1	*machineLearningTitanic*.ipynb
Data file for part 1	Titanic.csv
Data file for part 2	lowbtwt.csv
Dataset description for Part 2	lowbwt dataset description.docx

Very Important

The assignment due date cannot be changed. Late submission will incur 5% points earned penalty for each day late.

Do not upgrade the IBM clous account and do not enter the credit card number. This upgrade cannot be undone. You will be responsible for all incurred cost, and the faculty may not be able to assist with any technical issues.

Do not open any IBM cloud tickets unless instructed by the course TA and/or by the course instructor.

Do not deactivate the IBM cloud account. This action cannot be undone. Moreover, you will not be able to get another feature code until June.

Since we are working in the cloud environment, the updates are ongoing and the screenshots in instructions might not exactly match the interface. The substantial updates before the assignment due date will be handled within 24 hours. However, the instructions will not be updated past the assignment due date.

Assignment deliverables – 3 (three) files

- The HTML file saved in Part 1 (see details below)
- The HTML file saved in Part 2 (see details below)
- A single Word or PDF file with the writeup for Parts 1 and 2 and the answers to research questions in part 3. (see details below)

Only one assignment submission is allowed. Make sure to attach 3 files before you click submit button.

No credit will be given for Part 1 if the submission is missing the HTML file.
No credit will be given for Part 2 if the submission is missing the HTML file.

Part 1 – Run Logistic Regression on Titanic Data

Use your IBM cloud credentials to login to IBM Cloud Pak for data at https://dataplatform.cloud.ibm.com/login.

Add the Titanic Data

Click on Navigation Menu at the top left in Figure 1.

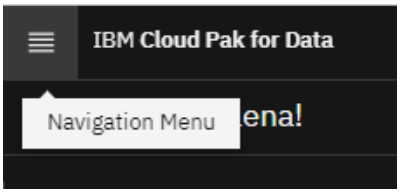
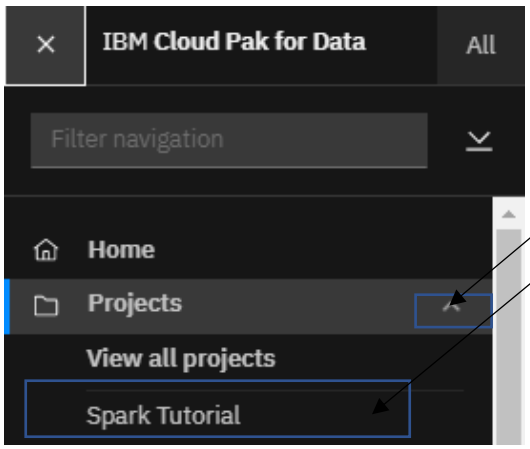


Figure 1: Navigation Menu

Choose Apache Spark Tutorial under Projects in Figure 2.

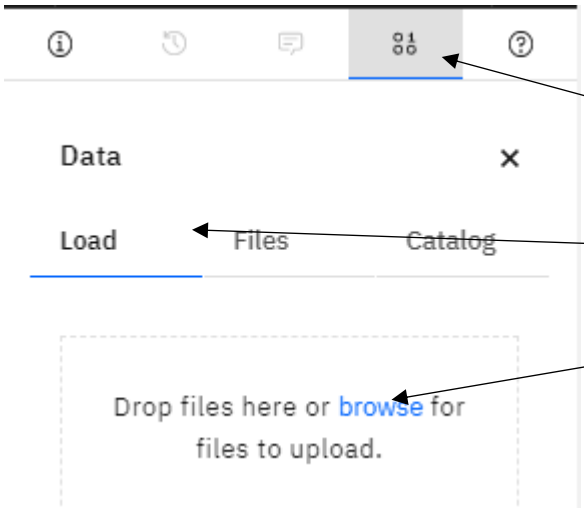


Click to Expand Projects

Choose Apache Spark Tutorial project

Note – Only recent projects appear in the menu. If you do not see your Apache Spark Tutorial project, select view all projects. The "my projects" page will open. Click on Apache Spark tutorial project name.

Figure 2: Choose Apache Spark Tutorial Project



Navigate to the Assets page of the project.

Click on a data icon on the top toolbar to open the slide out panel.

Make sure that Load tab is active.

Click on browse link and find the titanic.csv file that you downloaded from the course content.

Figure 3: Upload titanic.csv file

Add the Titanic Notebook

Click on Add to Project and Choose Notebook in Figure 4.

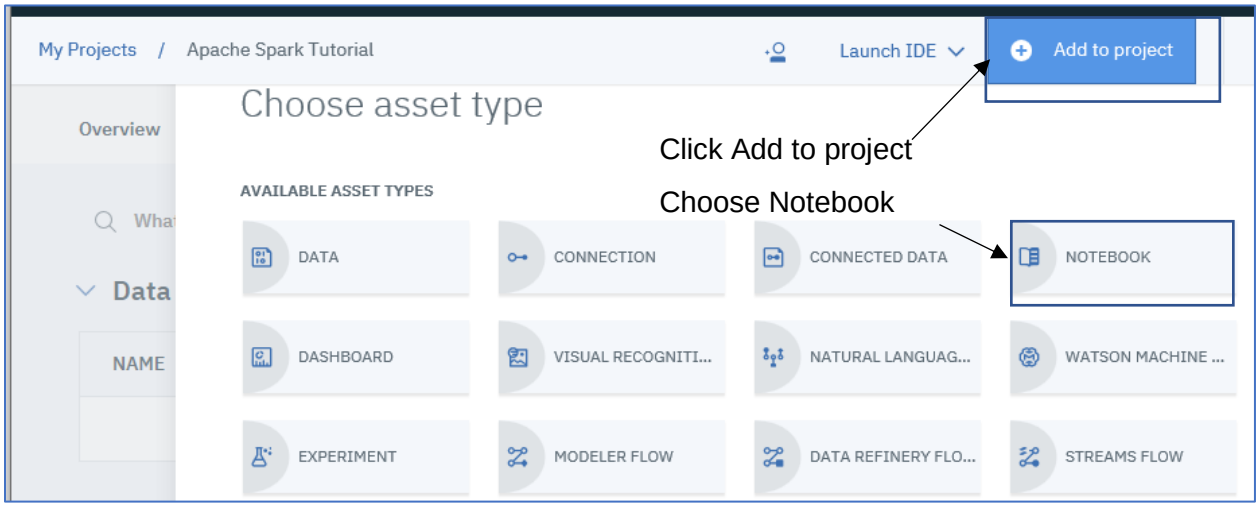


Figure 4: Add to project –notebook

Navigate to “from file” tab in Figure 5.

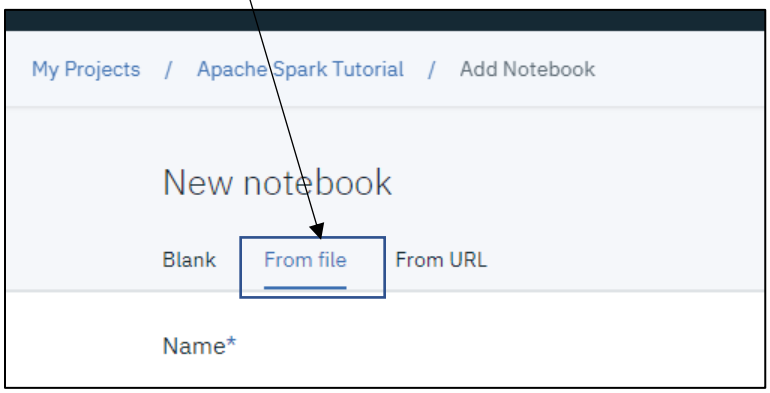


Figure 5: Choose From file

Scroll to Notebook file section in Figure 6. Upload machine learning Titanic notebook.

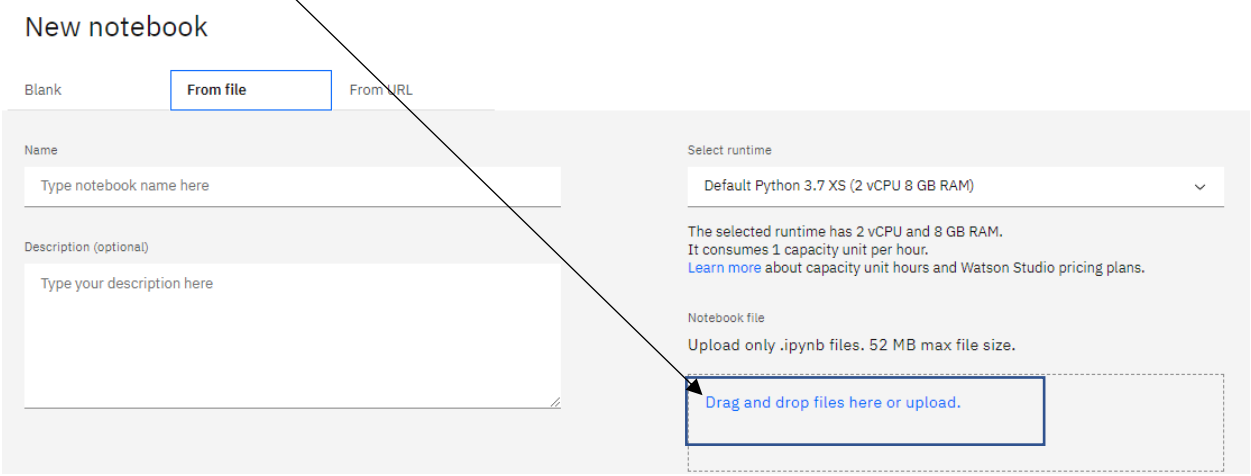


Figure 6: New Notebook – From file tab

The notebook name will be prepopulated with the file name but will remain editable.

Click on an arrow in the Select Runtime section. Select Default Spark 3.0 & Python 3.7 (driver:1 vCPU and 4 GB RAM, and 2 executors) option from the drop down.

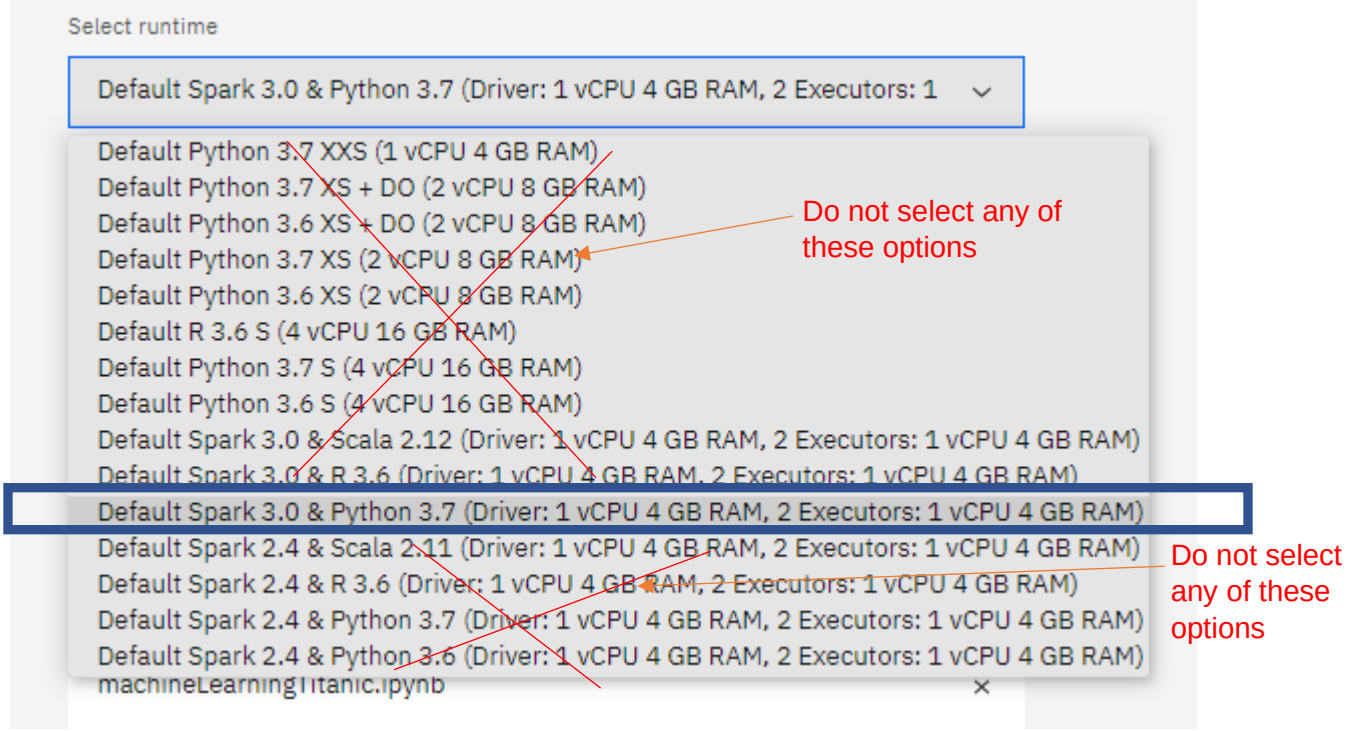


Figure 7: Choose the correct runtime – **do not skip this step**

Click Create to continue.

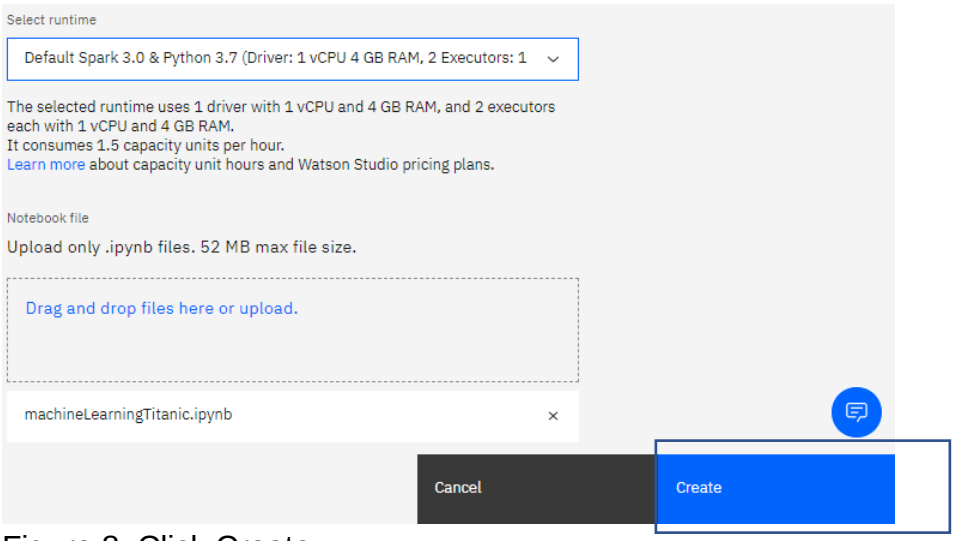


Figure 8: Click Create

The Spark Kernel will start in 1-2 minutes. You will see the notebook menu bar and the first two cells.

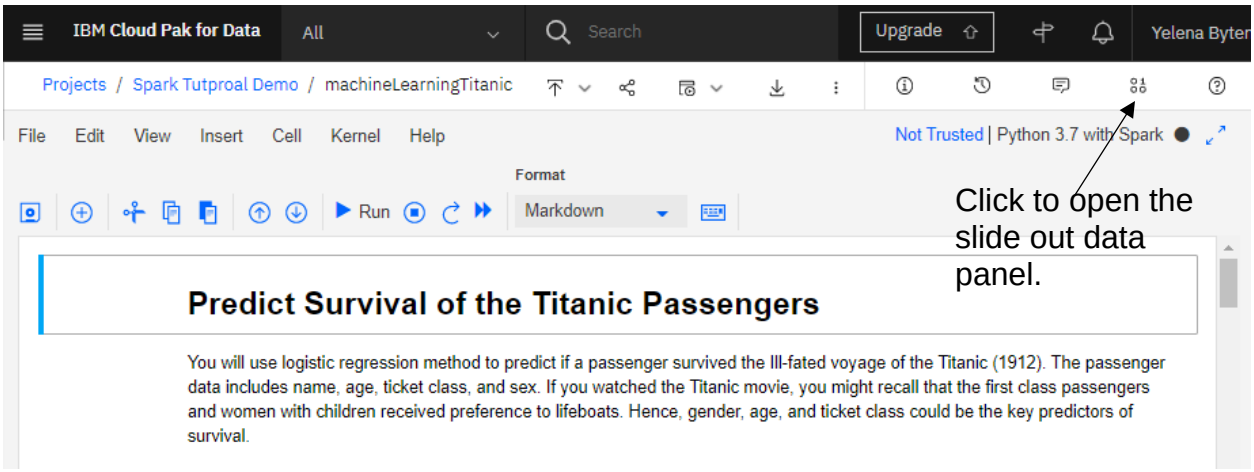


Figure 9: Notebook Interface and first two cells

Click on a data icon on the top toolbar to open the slide out panel.

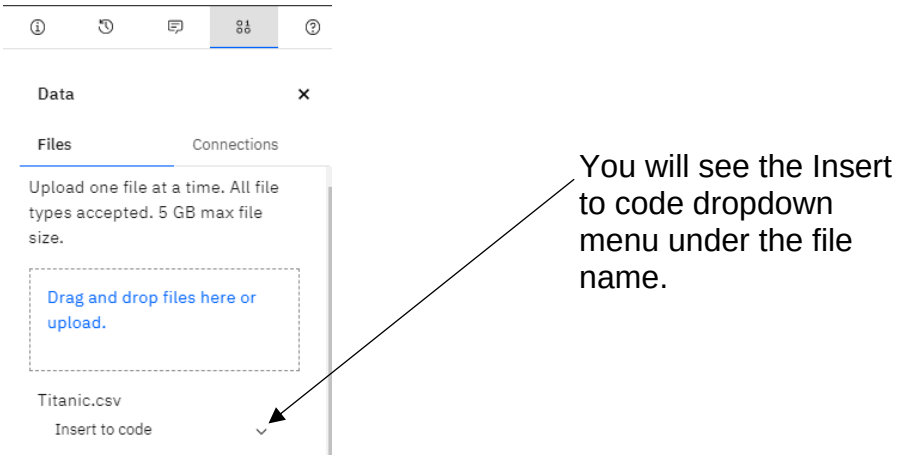


Figure 10: Data Slide-out Panel – Insert to Code options

Add Credentials Code

Scroll down to Access Data section. Place the cursor inside the executable cell and select Insert Spark Session DataFrame from the insert to code menu in the right panel.

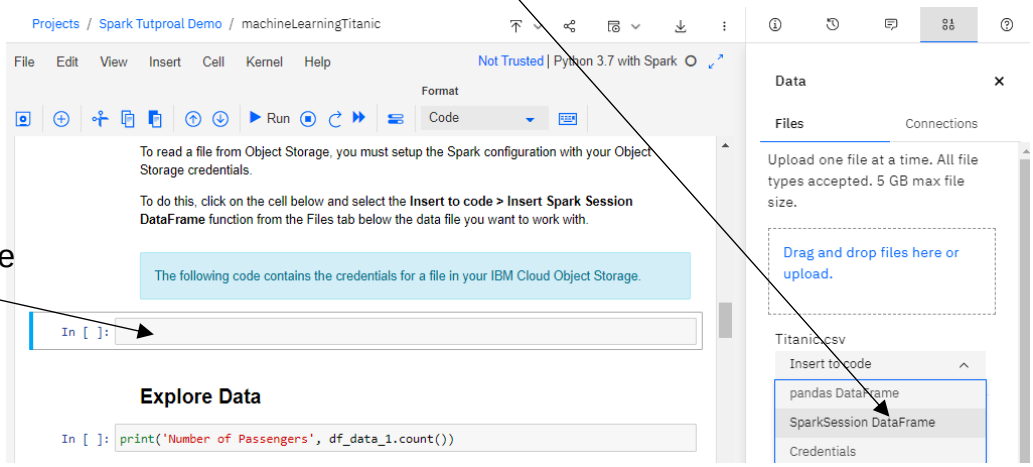


Figure 11: Insert Spark Session DataFrame

The last 7 lines of code load data into a Data Frame. Make sure that the Data Frame name is df_data_1. If not, then edit the code.

Add the following code under .option('header', 'true')\ to make sure that the columns have correct datatypes.

.option('inferSchema', 'true')\

Note that the object storage bucket name in Figure *12* contains the word “sparktutorial” because we are reusing the project from Spark Tutorial assignment,

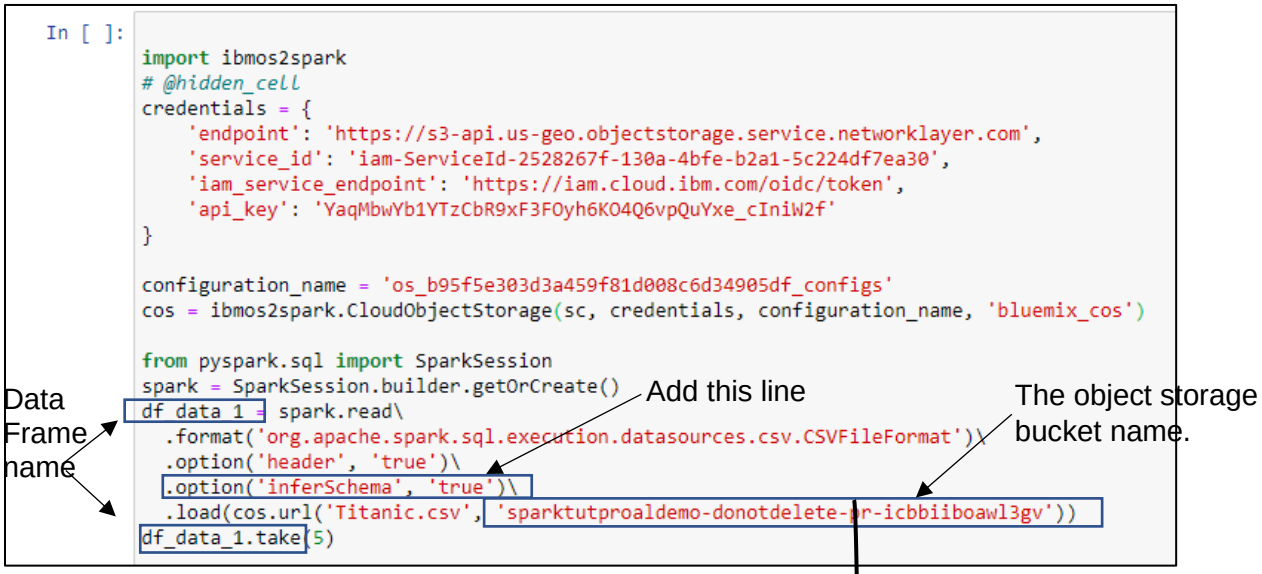


Figure 12: Credentials and code to load data

Update the Bucket Name

Scroll to the Parse Data section replace the string ‘BUCKET’ with the object storage bucket name.

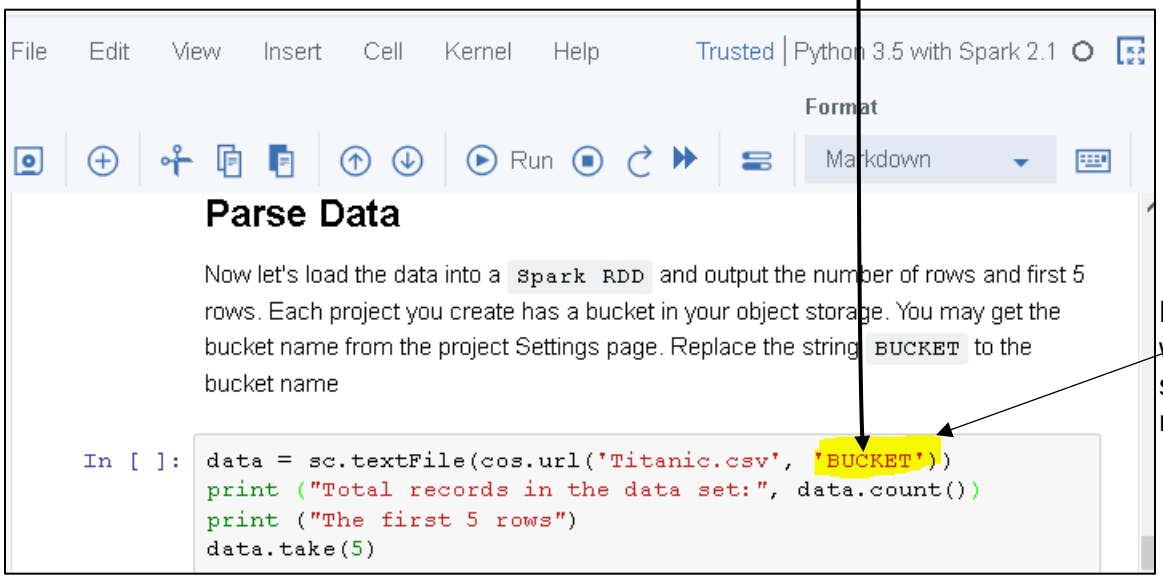


Figure 13: Replace BUCKET with Object Storage Bucket Name

Run all notebook cells starting from the top. Save the notebook with the output as an HTML document.

Writeup for part 1

Answer the following questions

- Discuss the importance of data exploration and visualization prior to running the logistic regression method.
- What do the Titanic data exploration results reveal about the relationships between the likelihood of survival and passenger data?
- Discuss the logistic regression method results, including the classification accuracy for training and test set.
- Is logistic regression suitable for this problem? Why or Why not?
- What alternative machine learning methods could be suitable for this problem? Consider at least 2 alternative methods.

Part 2 – Build Logistic Regression method for low birth weight data

Use the lowbtwt.csv data that you downloaded from Assignment 2 folder to build the logistic regression model in a new notebook.

Required Notebook Content

- Markdown cell with the dataset description at the top of the notebook
- Table of contents
- Code to load the necessary packages
- Credentials,
- Write a parsing function for the new data by carefully considering what the input variables should be and what the training target variable should be.
- Data exploration and visualization prior to running logistic regression. Include the histogram for the age variable, the frequency counts for the target variable, and the plots/frequency tables to check which variable could be a predictor for the target variable.
- Split the data into a training and test set
- Use the training set to build the logistic regression model.
- Tabulate the predicted and actual outcome
- Evaluate the model on the test set. Tabulate the predicted and actual outcome for the test set.
- The precision, recall, and F-measure for the training and test set.
- Roc curve plot
- Markdown cells with comments on what the code does and headings

Once the code works as expected, download the notebook as an HTML file.

Writeup for part 2

- Define the purpose of the study and the target variable. Which variables are used as predictors?
- Interpret the data exploration and visualization results. What did you learn about

the low birth weight data from data exploration, including possible relationships between predictors and the target variable?
- Discuss the method results, including the classification accuracy for training set and test set and model evaluation metrics (precision, recall, ROC curve area).
- Is the logistic regression method suitable for this study? Why or why not?
- How would you improve the accuracy of your model?
- Discuss at least 2 alternative machine learning methods that could be suitable for this problem and explain why?

Part 3 – Research questions

Answer the following questions in the paper. The answers must be your original work.

- What is overfitting? What is the impact of overfitting on model performance? Discuss at least 2 approaches to avoid overfitting the model.
- Discuss 5 (five) key differences between HDFS and Object Storage.
- We may use R, Python, Scala, and Java programming languages with Spark. Discuss the pros and cons of each language.

Assignment 2 Grading Rubric

The assignment deliverable must be an original work and must comply with the APA format and with the graduate standards. Cite all references you use to answer questions in Parts 1, 2, and 3.

Parts 1 and 2 cannot be graded if the HTML files are missing.

Criteria	Weight	Points Earned	Comments
Part 1 – Logistic regression on Titanic data (10 points for HTML submission and discussion; 20 points for answering questions) **No credit will be given for this part if the submission is missing the HTML file.** • Submitted the HTML file saved after running the notebook • The HTML file includes the cells output. • Data exploration and visualization discussion • Results interpretation and alternative methods discussion shows an application of Spark Machine Learning concepts.	30		

Part 2 – Logistic regression on low birth weight data. (15 points for HTML submission and discussion; 25 points for answering questions) **No credit will be given for this part if the submission is missing the HTML file.** • The submitted HTML file shows the working code and the code output • The notebook must include all required content. • The notebook includes markdown cells with comments, dataset description, and table of contents. • Clear definition of the study purpose and an identification of the target variable. • Data exploration and data visualization discussion • Method results interpretation, study limitations, and discussion of 2 alternative methods, including the rationale • Model evaluation metrics and an approach to improve the accuracy. • Demonstrate the concepts application through depth of analysis and findings discussion	40		
Part 3 – Respond to research questions (10 pts each) All answers should be in your own words. The references are used effectively to support the answers. The ideas from references are paraphrased. No word-to-word quotes. Depth of analysis. The answer must follow the APA format and should not have any spelling and grammatical errors.	30		
Total			