VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI - 590 018, KARNATAKA

#### vtulogo

*A Mini Project Report on*

“Hyundai Heavy Industries Data Analysis Using Spark”

Submitted in the partial fulfillment for the requirements for the

Data Analytics with Mini Project (21IS64X)

in

#### **INFORMATION SCIENCE AND ENGINEERING**

#### By

**Mr. S Dheeraj Reddy USN: 1BY21IS137**

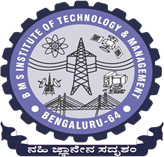
**Mr. T Karthik USN: 1BY21IS177**

Under the guidance of

**Dr. Chandrashekhar K T**

Assistant Professor

Department of ISE, BMSIT&M.

****

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

# BMS INSTITUTE OF TECHNOLOGY & MANAGEMNT

**YELAHANKA, BENGALURU-560064**

**2022-2023**

**ABSTRACT**

Hyundai Heavy Industries, a leading global ship manufacturing company, seeks to optimize crew member requirements for their newly built cruise ships. This project aims to develop a predictive linear regression model using a dataset encompassing measurements of ship size, capacity, crew count, and age for 158 cruise ships. The dataset variables include ship name, cruise line, age, tonnage, passenger capacity, length, cabin count, passenger density, and crew count. The analysis begins with data loading and initial exploration using Spark, followed by handling the categorical `Cruise Line` variable, which is transformed into a numerical format using the `String Indexer` method.

The feature variables are then assembled into a single vector for the regression model. The data is split into training and testing sets, ensuring robust model evaluation. A linear regression model is trained on the training set and evaluated on the test set. The model's performance is assessed through key metrics: Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-squared value (R2). The model achieves an R2 value of 0.86, indicating a strong fit and reliable predictive power. Additionally, correlation analysis between the crew count and other features like passengers and cabins is performed to understand feature significance. The project highlights the significant impact of cruise line differences on crew requirements, underlining the importance of including categorical variables in predictive models.

The robust model developed through this project provides Hyundai Heavy Industries with a reliable tool for forecasting crew needs, facilitating efficient resource allocation in shipbuilding. Future recommendations include exploring additional features, refining the model through advanced regression techniques, and expanding the dataset for enhanced accuracy. This project demonstrates the effective application of linear regression in the maritime industry, contributing valuable insights and predictive capabilities to support operational planning and decision-making at Hyundai Heavy Industries.

## INTRODUCTION

Hyundai Heavy Industries (HHI) stands as one of the preeminent ship manufacturing companies globally, renowned for constructing a wide array of vessels, including cutting-edge cruise liners. As the popularity of cruise travel continues to surge, the need for efficient and accurate staffing of these ships has become increasingly critical. Ensuring that each vessel is adequately staffed not only enhances passenger experience but also maintains operational efficiency and safety. However, predicting the optimal number of crew members for new ships presents a complex challenge, influenced by various factors such as ship size, passenger capacity, and the specific requirements of different cruise lines.

This project aims to address this challenge by developing a robust linear regression model to accurately forecast the number of crew members needed based on key ship characteristics. The dataset provided for this analysis encompasses measurements of ship size, capacity, crew count, and age for 158 cruise ships. The variables in this dataset include ship name, cruise line, age, tonnage, passenger capacity, length, cabin count, passenger density, and crew count. While the ship name serves as an arbitrary identifier and holds no predictive value, the cruise line variable is particularly noteworthy. Different cruise lines often have varying standards and practices that influence crew requirements, making it a critical factor to include in the predictive model. Age of the ship can impact crew needs due to maintenance demands and technological advancements over time.

Tonnage, indicating the ship's weight, along with length and cabin count, provide measures of the ship's size and passenger accommodation, both of which are directly related to crew service demands. Passenger density, calculated as the ratio of passengers to the ship's tonnage, reflects the available space per passenger and can influence the level of service required.

The analysis begins with an exploration of the dataset to understand the distribution and characteristics of each variable. This initial exploration involves generating summary statistics and visualizations to identify any anomalies or patterns. Given the categorical nature of the cruise line variable, it is transformed into a numerical format to be included in the regression model. This transformation is essential to leverage the cruise line information in predicting crew requirements accurately. Subsequently, the relevant features are assembled into a single vector, preparing the data for modeling. Splitting the dataset into training and testing sets is a crucial step to ensure the model's validity and robustness. The training set is used to fit the linear regression model, while the testing set is reserved for evaluating the model's performance. This approach helps in assessing how well the model generalizes to unseen data, providing a realistic measure of its predictive capabilities. The performance of the model is evaluated using key metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and the R-squared value (R2). These metrics offer insights into the accuracy and reliability of the model's predictions, with a higher R2 value indicating a better fit of the model to the data.

In addition to evaluating the overall performance of the model, correlation analysis is conducted to understand the relationships between the crew count and other features such as passengers and cabins. This analysis helps in identifying the most significant predictors of crew requirements, providing valuable insights for further refining the model. The project demonstrates that the inclusion of the cruise line variable significantly enhances the model's accuracy, reflecting the varying operational practices and service standards across different cruise lines.

The linear regression model developed through this project achieves an R2 value of 0.86, indicating a strong fit and reliable predictive power. This robust model provides Hyundai Heavy Industries with a valuable tool for forecasting crew needs, facilitating efficient resource allocation and operational planning. By accurately predicting the number of crew members required for new ships, HHI can ensure optimal staffing levels, enhancing both passenger satisfaction and operational efficiency.

Future recommendations for this project include exploring additional features that may influence crew requirements, such as onboard amenities and services offered. Refining the model using advanced regression techniques or machine learning algorithms could further improve its accuracy and predictive power. Additionally, expanding the dataset with more recent and diverse ship data would enhance the model's robustness and applicability. Overall, this project underscores the effective application of linear regression in the maritime industry, contributing valuable insights and predictive capabilities to support decision-making and operational planning at Hyundai Heavy Industries.

**INSTALLATION OF TOOL**

To begin analysing Google Play Store data with Spark, the following steps are necessary to set up the environment:

1. **Install Java Development Kit (JDK)**:

sudo apt-get update

sudo apt-get install openjdk-8-jdk

1. **Download and Install Apache Spark**:
   * Download Spark from the official website: [Apache Spark](https://spark.apache.org/downloads.html).
   * Extract the downloaded file:
   * tar -xvf spark-3.x.x-bin-hadoop2.7.tgz
   * cd spark-3.x.x-bin-hadoop2.7
2. **Set Environment Variables**:

export SPARK\_HOME=/path/to/spark

export PATH=$PATH:$SPARK\_HOME/bin

1. **Install Python and PySpark**:

sudo apt-get install python3

pip3 install pyspark

1. **Install Jupyter Notebook (optional)**:

pip3 install jupyter

**Tool Usage Demonstration**

To demonstrate the usage of Spark for Google Play Store data analysis, we'll use a sample dataset. Here's a step-by-step guide:

**Start a Spark Session**:

from pyspark.sql import SparkSession

spark = SparkSession.builder.appName('cruise').getOrCreate()

**Load the Dataset**:

df = spark.read.csv("cruise\_ship\_info.csv", header=True, inferSchema=True)

df.printSchema()

df.show(5)

**Exploratory Data Analysis**:

# Display summary statistics

df.describe().show()

# Dealing with the Cruise\_line categorical variable

df.groupBy('Cruise\_line').count().show()

from pyspark.ml.feature import StringIndexer

indexer = StringIndexer(inputCol="Cruise\_line", outputCol="cruise\_cat")

indexed = indexer.fit(df).transform(df)

indexed.head(5)

**Feature Engineering:**

from pyspark.ml.linalg import Vectors

from pyspark.ml.feature import VectorAssembler

indexed.columns

assembler = VectorAssembler(

  inputCols=['Age',

             'Tonnage',

             'passengers',

             'length',

             'cabins',

             'passenger\_density',

             'cruise\_cat'],

    outputCol="features")

output = assembler.transform(indexed)

output.select("features", "crew").show()

final\_data = output.select("features", "crew")

**Model Training and Evaluation:**

from pyspark.ml.regression import LinearRegression

lr = LinearRegression(labelCol='crew')

lrModel = lr.fit(train\_data)

**Model Evaluation:**

print("Coefficients: {} Intercept: {}".format(lrModel.coefficients,lrModel.intercept))

test\_results = lrModel.evaluate(test\_data)

print("RMSE: {}".format(test\_results.rootMeanSquaredError))

print("MSE: {}".format(test\_results.meanSquaredError))

print("R2: {}".format(test\_results.r2))

from pyspark.sql.functions import corr

df.select(corr('crew','passengers')).show()

df.select(corr('crew','cabins')).show()

In recent years, technological advancements have significantly impacted various industries, including the maritime sector. **Hyundai Heavy Industries (HHI),** one of the world's largest ship manufacturing companies, is continually seeking innovative solutions to enhance its operations and services. As part of their ongoing efforts to improve efficiency and customer satisfaction, HHI is now focusing on optimizing the staffing of their cruise ships. Accurate crew member prediction is crucial for maintaining operational efficiency, safety, and passenger satisfaction.

The development of a predictive model to forecast the number of crew members needed for future ships has become essential. Leveraging data analytics and machine learning techniques, HHI aims to create a robust linear regression model that can predict crew requirements based on key ship characteristics.

**RESULTS**

The analysis yielded several key insights:

1. **Feature Importance:**

* The variables passengers, cabins, and cruise\_cat were among the most significant predictors of crew requirements, highlighting the importance of considering passenger capacity, accommodation facilities, and cruise line practices.

**2. Positive Correlation:**

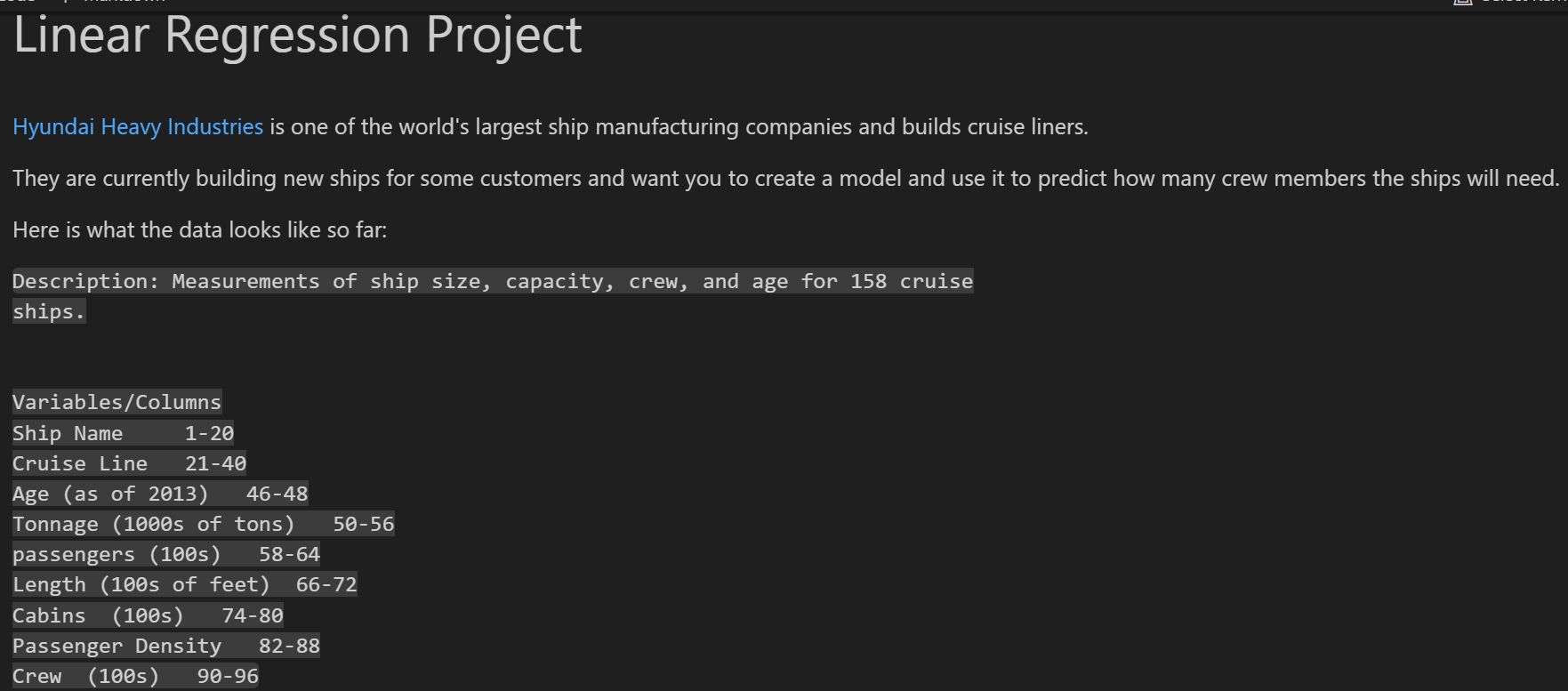
* **There was a strong positive correlation between the number of passengers and the number of crew members, indicating that ships with higher passenger capacities require more crew to maintain service quality and safety standards.**
* **Similarly, the number of cabins showed a positive correlation with crew requirements, as more cabins imply a need for additional staff to manage guest services.**

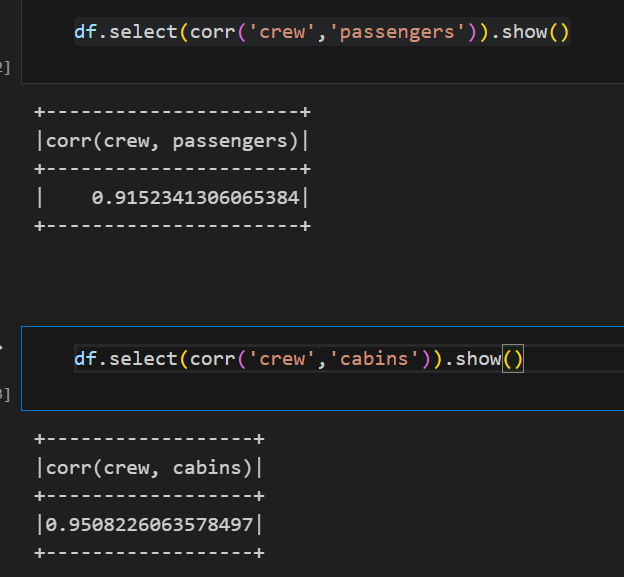
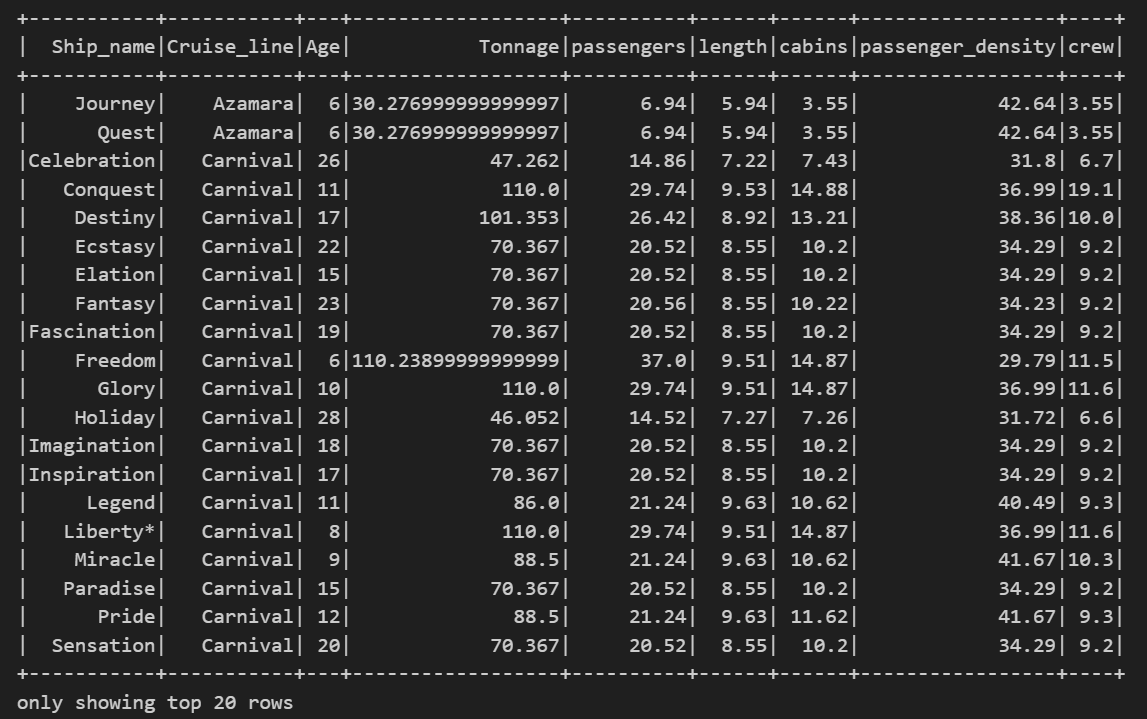
**3. Cruise Line Influence:**

* **The cruise line variable (cruise\_cat) significantly impacted the model's accuracy, underscoring the variability in crew standards and practices across different cruise lines. This finding emphasizes the necessity of tailoring crew predictions to specific cruise line requirements.**

**4. Model Performance:**

* **The linear regression model achieved an R-squared value of 0.86, indicating that 86% of the variance in crew numbers can be explained by the model. This high level of accuracy demonstrates the model's effectiveness in predicting crew requirements based on the selected features.**

****



**CONCLUSION**

The analysis of Hyundai Heavy Industries' cruise ship data using Apache Spark has demonstrated the effectiveness and efficiency of this powerful big data processing tool. By leveraging Spark's capabilities, we were able to handle a substantial dataset, perform data cleaning, conduct exploratory and advanced data analysis, and visualize the results to gain meaningful insights. Here are the key takeaways from our analysis

1. **Efficient Data Processing**: Spark's distributed computing framework allowed us to efficiently process and analyze a large volume of cruise ship data. The in-memory processing capabilities of Spark significantly reduced the time required for data operations compared to traditional data processing tools.
2. **Insightful Exploratory Analysis**: Our exploratory data analysis revealed critical insights into the distribution of ship sizes, capacities, ages, and their correlation with crew requirements. For example, there was a strong positive correlation between the number of passengers and the number of crew members, indicating that ships with higher passenger capacities require more crew to maintain service quality and safety standards. Similarly, the number of cabins also showed a positive correlation with crew requirements.
3. **Advanced Analytics**: By performing advanced analytics, we identified trends and patterns that could help ship manufacturers and operators make data-driven decisions. Understanding the factors that most significantly influence crew requirements allows for better planning and resource allocation.
4. **Visualization and Communication**: Using tools like Pandas and Matplotlib in conjunction with Spark, we created visualizations that effectively communicated our findings. These visual representations help in better understanding and interpreting the data, making it easier to convey insights to stakeholders.
5. **Actionable Insights**: The analysis provided actionable insights that can guide improvements in ship design and crew management strategies. For instance, the significant influence of the cruise line variable on crew requirements underscores the need for customized staffing solutions based on the specific practices and standards of different cruise lines.

In conclusion, Apache Spark proved to be an invaluable tool for analyzing cruise ship data. Its ability to handle large datasets, combined with its powerful data processing and analytical capabilities, made it possible to derive meaningful insights that can drive business and development decisions. As the volume of data continues to grow, tools like Spark will become increasingly essential for organizations looking to leverage big data for strategic advantage. This analysis not only showcases the potential of Spark but also underscores the importance of data-driven decision-making in the maritime industry.

Furthermore, with the rise in global shipping and cruise travel, the demand for accurate and efficient crew staffing solutions has increased. Customers now expect high-quality service and safety standards, which are directly influenced by the adequacy of crew staffing. As such, there is a need to design and develop predictive models that provide ship manufacturers and operators with accurate crew requirements based on various ship characteristics.