**Phase 2: Future Sales Prediction**

**Abstract:**

Future sales prediction entails the sophisticated utilization of advanced analytical methodologies, notably regression analysis and machine learning paradigms, to prognosticate prospective sales trends predicated on antecedent historical data. It encompasses the discerning extraction of intricate patterns, correlations, and discernments immanent within the dataset to formulate judicious estimations concerning the quantum and typology of products or services that an enterprise is poised to transact in the ensuing temporal domain. This proactive modus operandi confers upon organizations the capacity to finesse inventory dynamics, meticulously calibrate resource allocation, and refine overall operational efficacy, thereby cultivating a discernible competitive advantage within the intricate tapestry of the commercial milieu.

In the dynamic landscape of modern business, the ability to anticipate and adapt to future sales trends is paramount for sustained success. This project explores the application of regression analysis and machine learning techniques to predict future sales. By leveraging historical sales data and relevant features, our objective is to develop a robust predictive model that provides actionable insights for decision-makers.

**2.1 Problem Description:**

The problem is to develop a predictive model that uses historical sales data to forecast future sales for a retail company. The objective is to create a tool that enables the company to optimize inventory management and make informed business decisions based on data driven sales predictions. This project involves data preprocessing, feature engineering, model selection, training, and evaluation.

“We are asking you to predict total sales for every product and store in the next month. By solving this competition you will be able to apply and enhance your data science skills.”

**2.2 Dataset Information**

We collected this dataset from Kaggle (<https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales/data>) . We also download it using

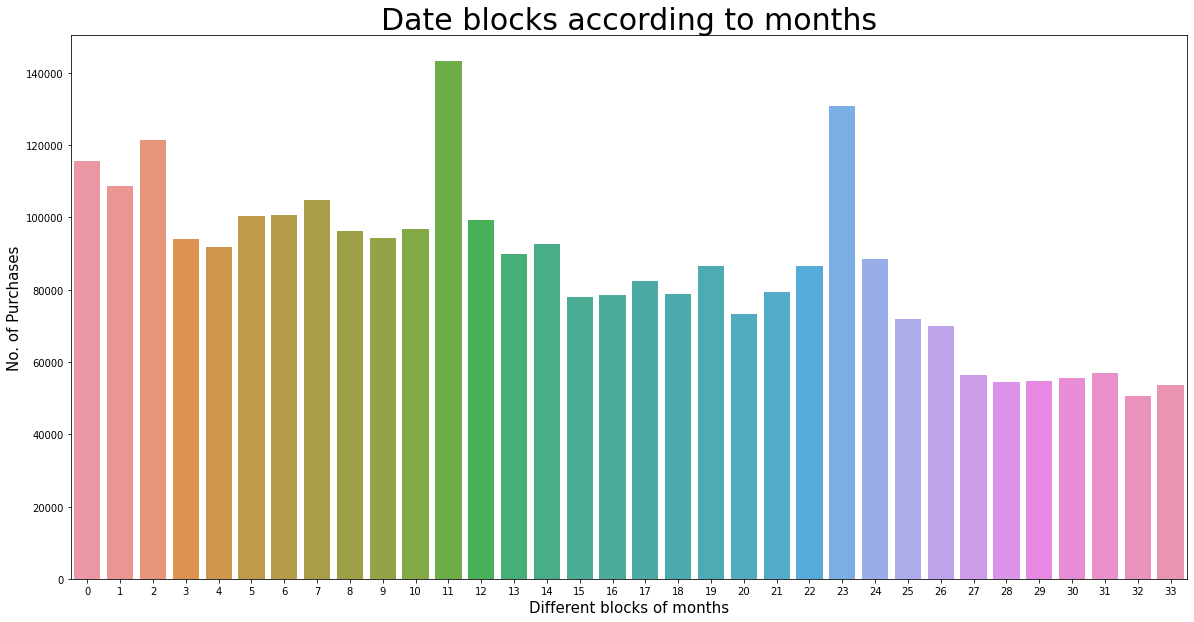
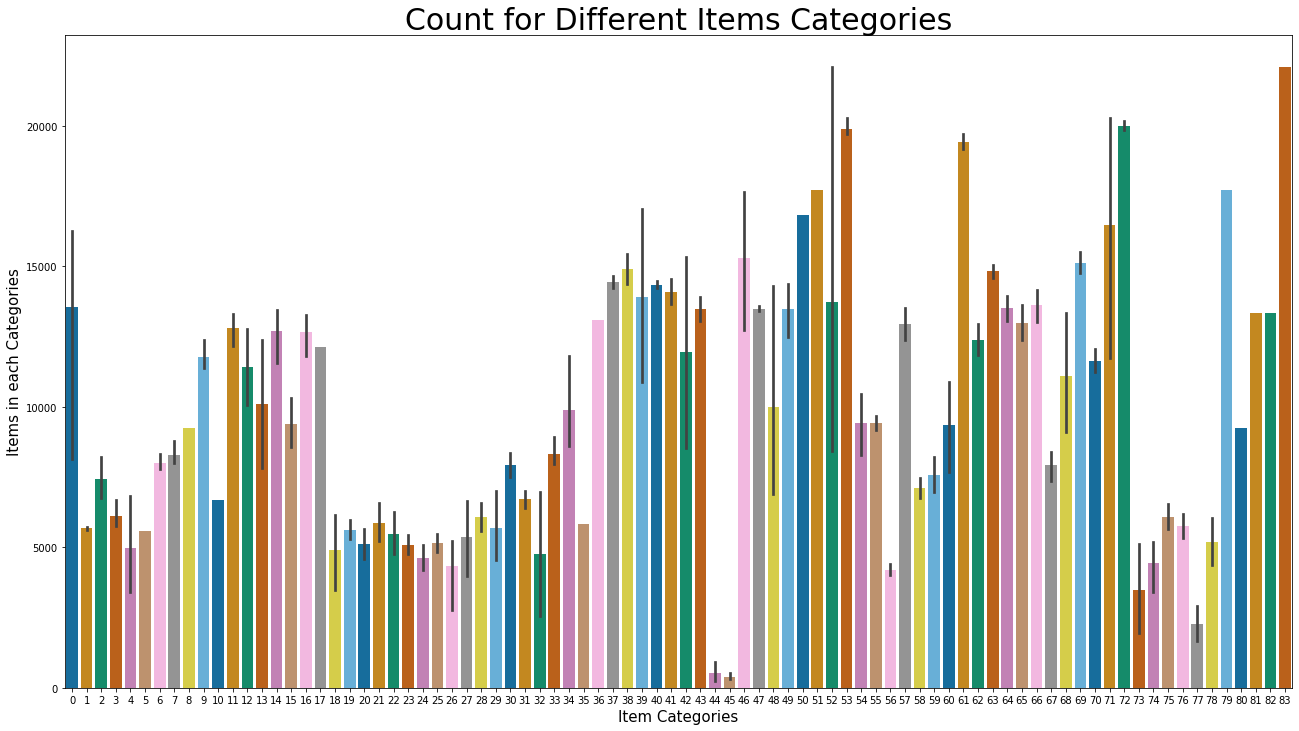
**> kaggle competitions download -c competitive-data-science-predict-future-sales**

**2.3.1 Dataset Columns**

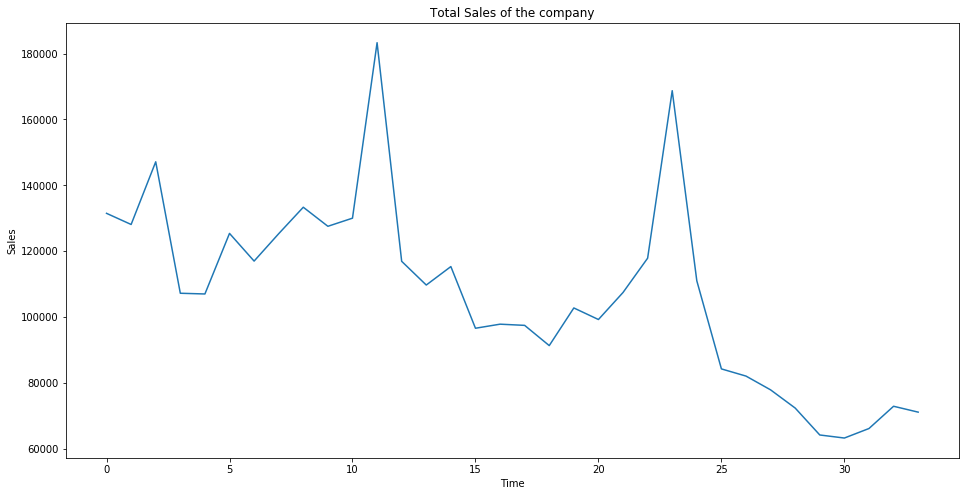
* **ID** : an Id that represents a (Shop, Item) tuple within the test set
* **shop\_id** : unique identifier of a shop
* **item\_id** : unique identifier of a product
* **item\_category\_id** : unique identifier of item category
* **item\_cnt\_day** : number of products sold. You are predicting a monthly amount of this measure
* **item\_price** : current price of an item
* **date** : date in format dd/mm/yyyy
* **date\_block\_num** : a consecutive month number, used for convenience.
* **item\_name** : name of item
* **shop\_name** : name of shop
* **item\_category\_name** : name of item category

Fig.1. Plot of Item categories and the items in each categories

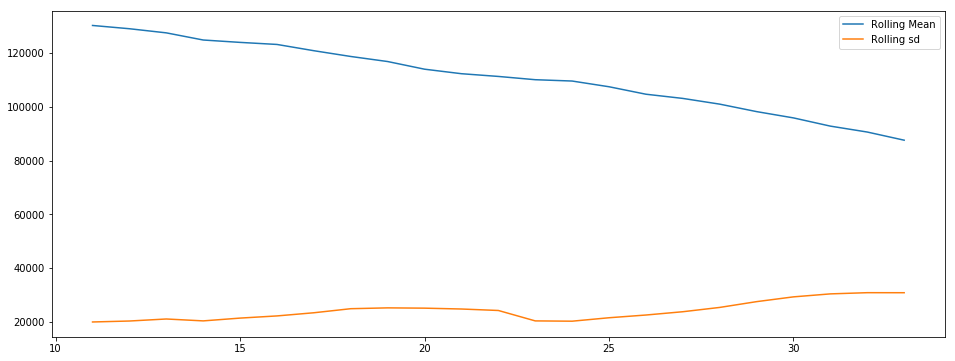
**2.3.2 Exploring the Dataset:**



Analyzing the Seasonality:



Determining the Trends:



**Quick observations:** There is an obvious "seasonality" (Eg: peak sales around a time of year) and a decreasing "Trend".

**2.4.1 Models/Libraries Used:**

**NumPy :**

**Purpose:**

NumPy is a powerful numerical computing library in Python. It provides support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays.

**Key Functions:**

* Efficient array operations and manipulation.
* Mathematical functions for array operations.
* Random number generation.

**Pandas :**

**Purpose:**

Pandas is a data manipulation and analysis library in Python. It provides data structures like Series and DataFrame for handling and analyzing structured data.

**Key Functions:**

* Reading and writing various data formats (CSV, Excel, SQL, etc.).
* Data cleaning, manipulation, and transformation.
* Handling missing data and merging datasets.

**Matplotlib :**

**Purpose:**

Matplotlib is a popular 2D plotting library for creating static, interactive, and animated visualizations in Python.

**Key Functions:**

* Creating line plots, scatter plots, bar plots, histograms, etc.
* Customizing and styling plots.
* Visualizing data distributions and trends.

**Seaborn:**

**Purpose:**

Seaborn is a statistical data visualization library based on Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics.

**Key Functions:**

* Simplifies the creation of complex visualizations.
* Enhances Matplotlib plots with additional styles and color palettes.
* Visualizing statistical relationships in data.

**2.4.2 Models Description:**

**Linear Regression:**

Linear Regression is a fundamental and widely used supervised learning algorithm for predicting a continuous outcome. The model assumes a linear relationship between the input features and the target variable. It works by finding the best-fit line through the data, minimizing the sum of squared differences between the predicted and actual values. Linear Regression is interpretable, and its coefficients provide insights into the impact of each feature on the target variable. While simple, it is effective when the underlying relationships in the data are approximately linear.

**Lasso Regression:**

Lasso Regression, or Least Absolute Shrinkage and Selection Operator, is a regularization technique applied to linear regression. It introduces a penalty term to the linear regression objective function, which is the sum of the absolute values of the coefficients multiplied by a regularization parameter (alpha). This penalty encourages sparsity in the model by driving some coefficients to exactly zero. Lasso Regression is particularly useful when dealing with high-dimensional data and can serve as a feature selection method, effectively ignoring less relevant features.

**Ridge Regression:**

Ridge Regression is another regularization technique for linear regression. Similar to Lasso, it introduces a penalty term to the linear regression objective function, but this time it is the sum of the squared values of the coefficients multiplied by a regularization parameter (alpha). Ridge Regression is effective in preventing overfitting, especially when there are correlated features in the dataset. It tends to shrink the coefficients towards zero, but unlike Lasso, it rarely sets them exactly to zero. Ridge is advantageous when retaining all features is desirable but with a controlled impact on their magnitudes.

In summary, Linear Regression forms the foundation, while Lasso and Ridge Regressions provide regularization mechanisms to handle different challenges such as multicollinearity and feature selection. The choice between these models depends on the characteristics of your dataset and the specific goals of your project.

**Training and Testing:**

Explanation:

Initialization:

Import the necessary modules, including LinearRegression from scikit-learn for creating the Linear Regression model.

Model Training:

* Initialize the Linear Regression model.
* Train the model on the training dataset (x\_train and y\_train) using the fit method.

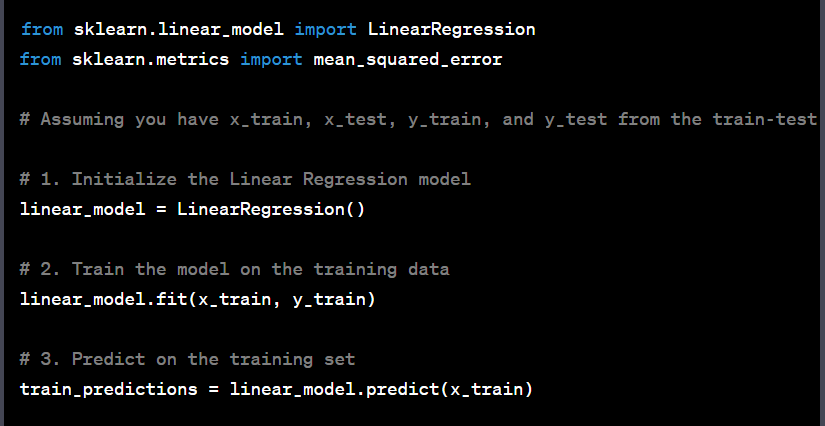
Training Prediction and Evaluation (Optional):

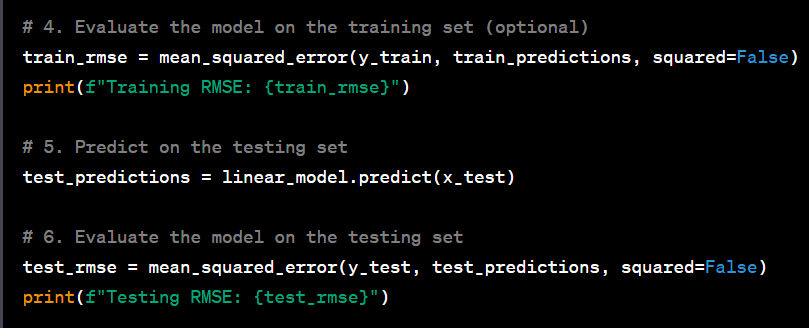
* Predict the target variable on the training set using the trained model.
* Optionally, calculate and print the Root Mean Squared Error (RMSE) to evaluate the model's performance on the training data.

Testing Prediction and Evaluation:

* Predict the target variable on the testing set using the trained model.
* Calculate and print the RMSE to evaluate the model's performance on the testing data.

Source Code:





**2.6 Data Correlation:**

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.