MARMARA UNIVERSITY

FACULTY of ENGINEERING

CSE4062 – Data Science Project Delivery #1

Spring, Group #3

"Machine Learning Applications on Issue Tickets"

Group Members

150117064 – Yunus Emre ERTUNC – yeertunc63@gmail.com
150117036 – Muhammed Enes AKTURK – enesakturk999@hotmail.com
150317048 – Sude SAHIN – sudesahin17@hotmail.com
150717039 – Musab AKICI – musabakici8@gmail.com
150418851 – Mehmet Emre OZ – mehmetemreoz35@gmail.com

Supervised by

Assoc. Prof. Dr. Murat Can
Ganiz

27.04.2022

Introduction

The main aim of this project is understanding the machine learning algorithms like classification, regression and clustering by working on huge amounts of data. Python programming language will be used during the implementation of these algorithms. Powerful libraries like Scikit-Learn and PyTorch will make the implementation of algorithms easier.

Overview

Machine learning is a branch of the artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn. Classification, regression and clustering are some tasks that machine learning concerns. Classification refers to a predictive modeling problem where a class label is predicted for a given example of input data, regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome and lastly clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

Project Definition

In this project we are given a dataset which is called as "Issue Tickets" and we are expected to perform several machine learning tasks on this dataset. There will be different deliveries of this project. At the beginning, we will examine the dataset given us in order to extract features and understand dataset better. Then, ML algorithms will be performed to get deeper understanding of the dataset.

Data Statistics

Dataset Name: Issue Tickets

Dataset File Type: .ods

ISSUE_ID field was removed before our analysis.

- Number of rows and number of columns of dataset:

# of Rows	# of Columns
16969	15

- The important features and their types are:

Feature	Type
REPORTER	Text
ISSUE_TYPE	Text
PRIORITY	Text
COMPNAME	Text
WORKER	Text
EMPLOYEE_TYPE	Text
WORK_LOG	Numeric
ISSUE_CATEGORY	Text
LABEL	Numeric
CREATION_DATE	Text
RESOLUTION_DATE	Text

- These are the target attributes that can be used for classification / regression:

Target Attributes	Classification / Regression
WORK_LOG	Regression
ISSUE_CATEGORY / LABEL	Classification

- Count of issues opened by each reporter:

Reporters VS Number of issues opened

Reporter Count : 378
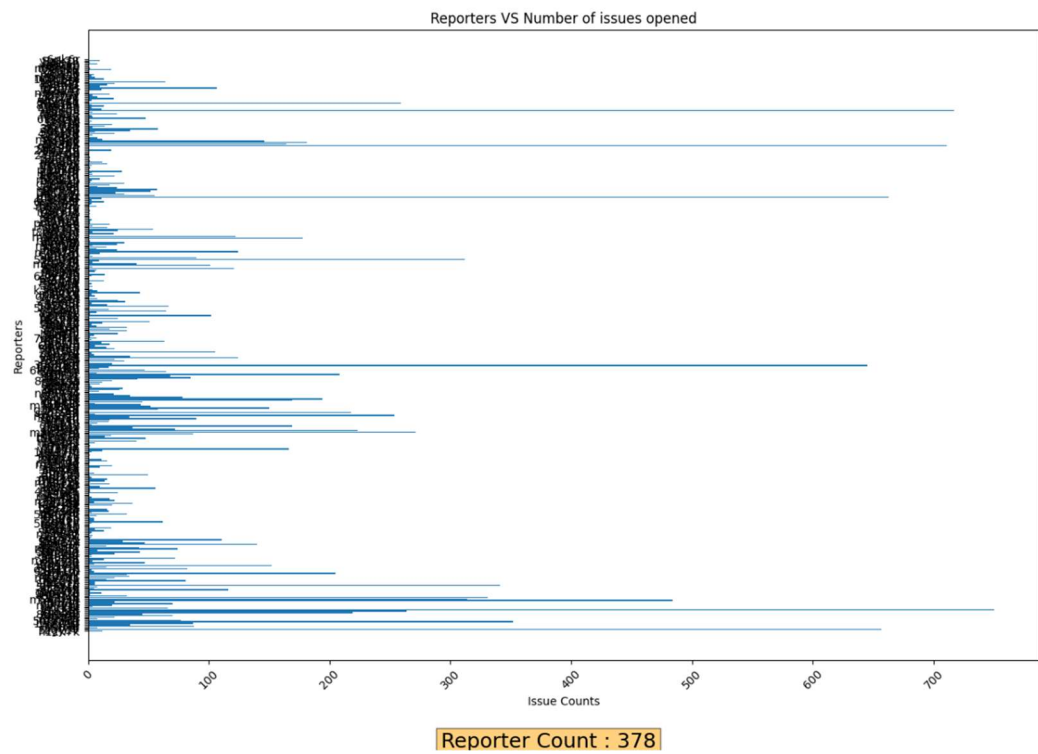
- Number of issue types and count of each issue type:

Issue Type Counts

Issue Types Count : 14
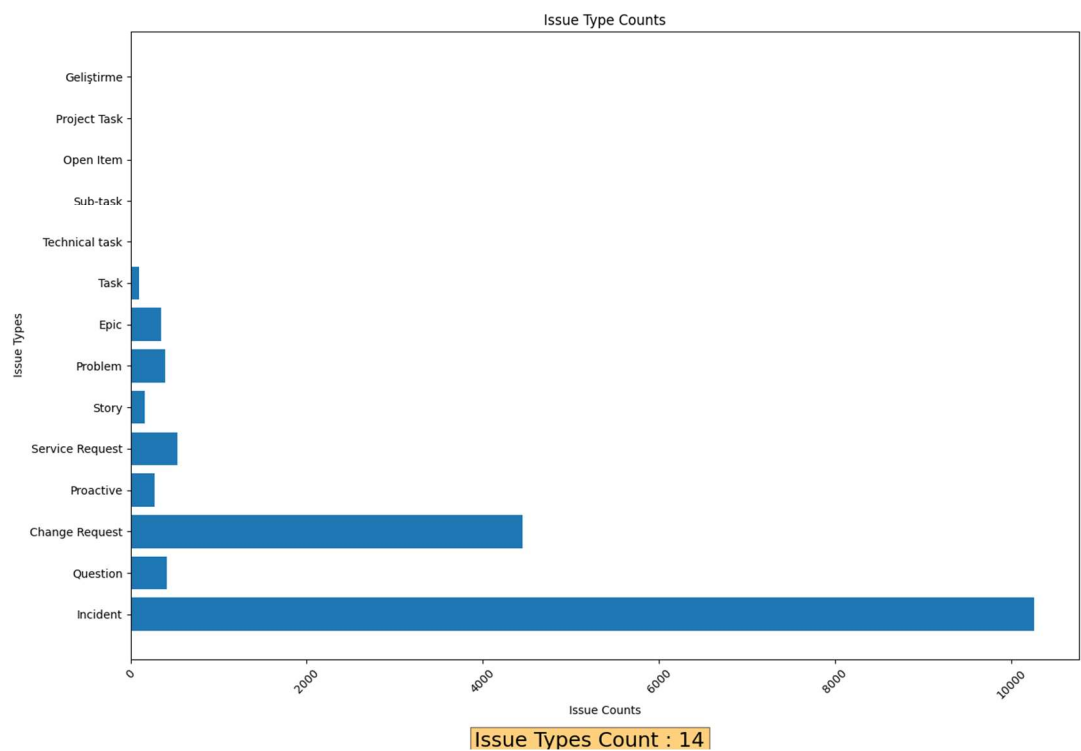
- Issue types and priority distribution:

Issue Type - Priority Count : 38
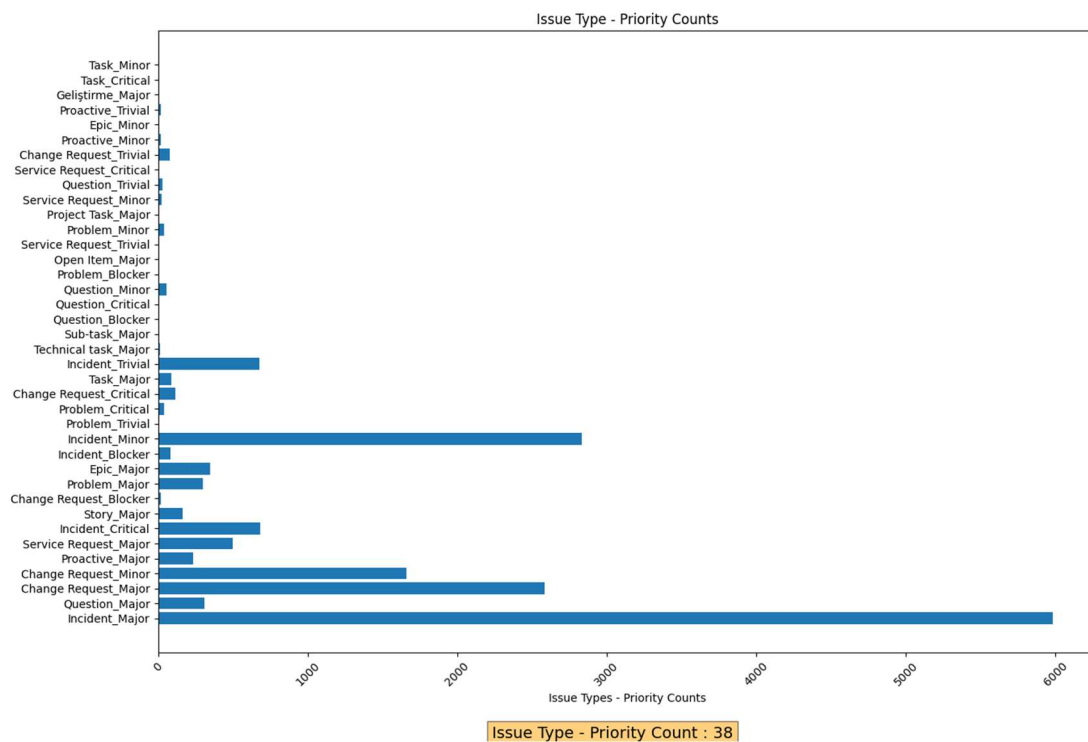
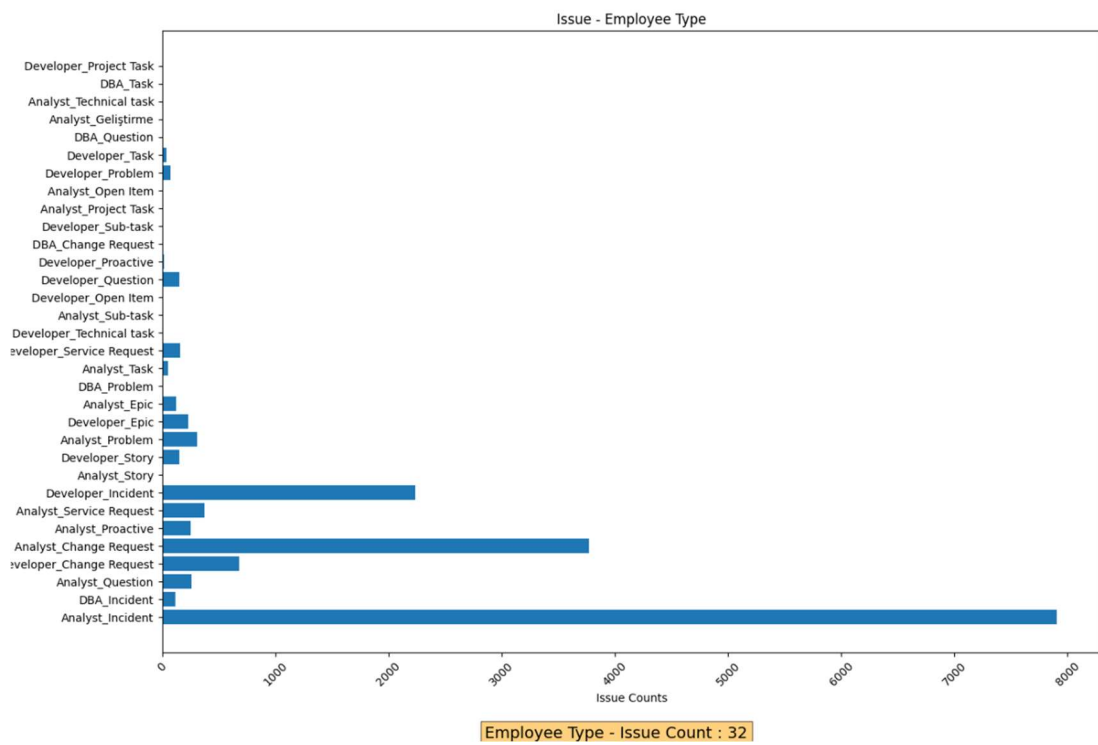- Number of issues opened by each company:

Company Count : 66
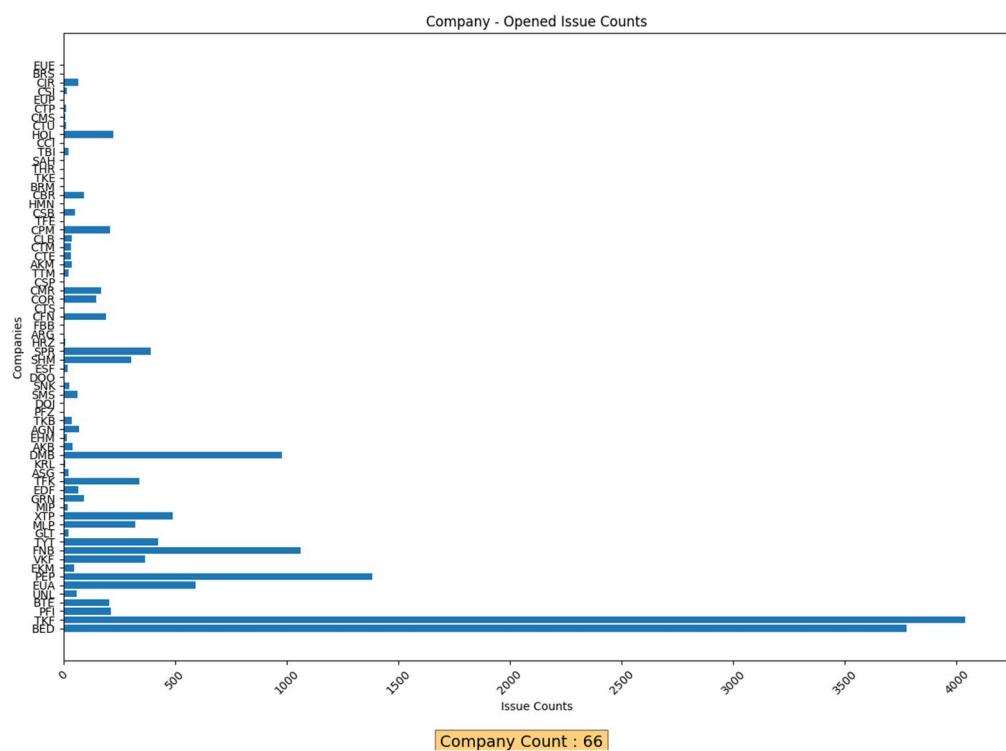
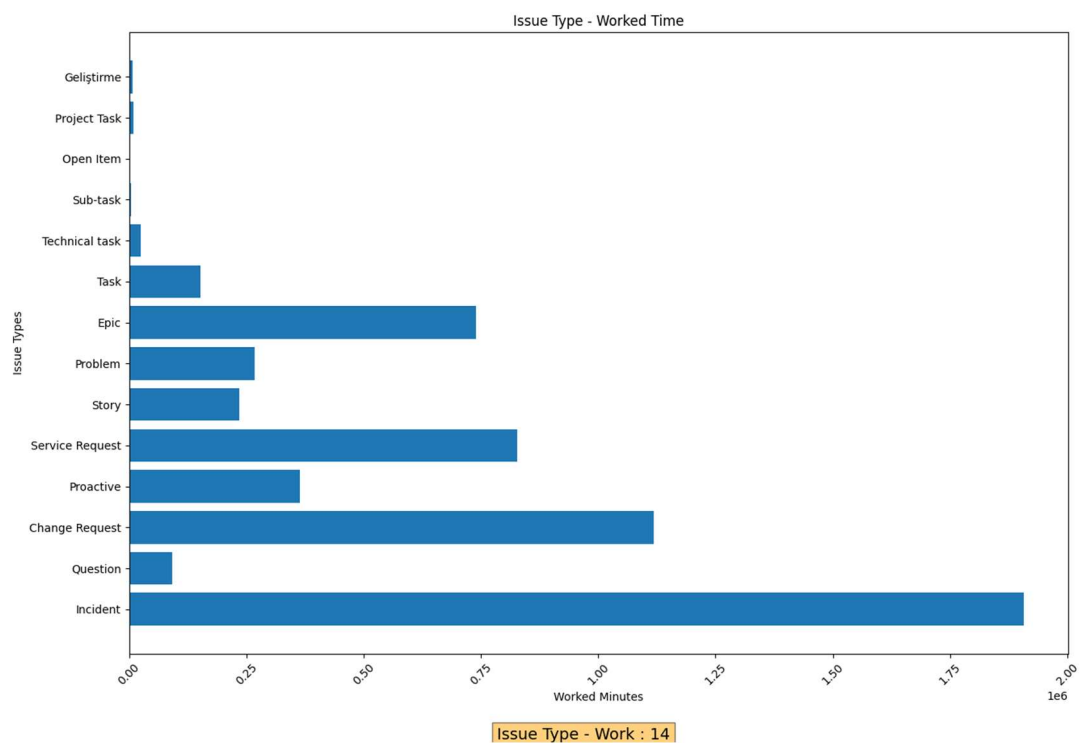- Number of issues that each worker worked on:
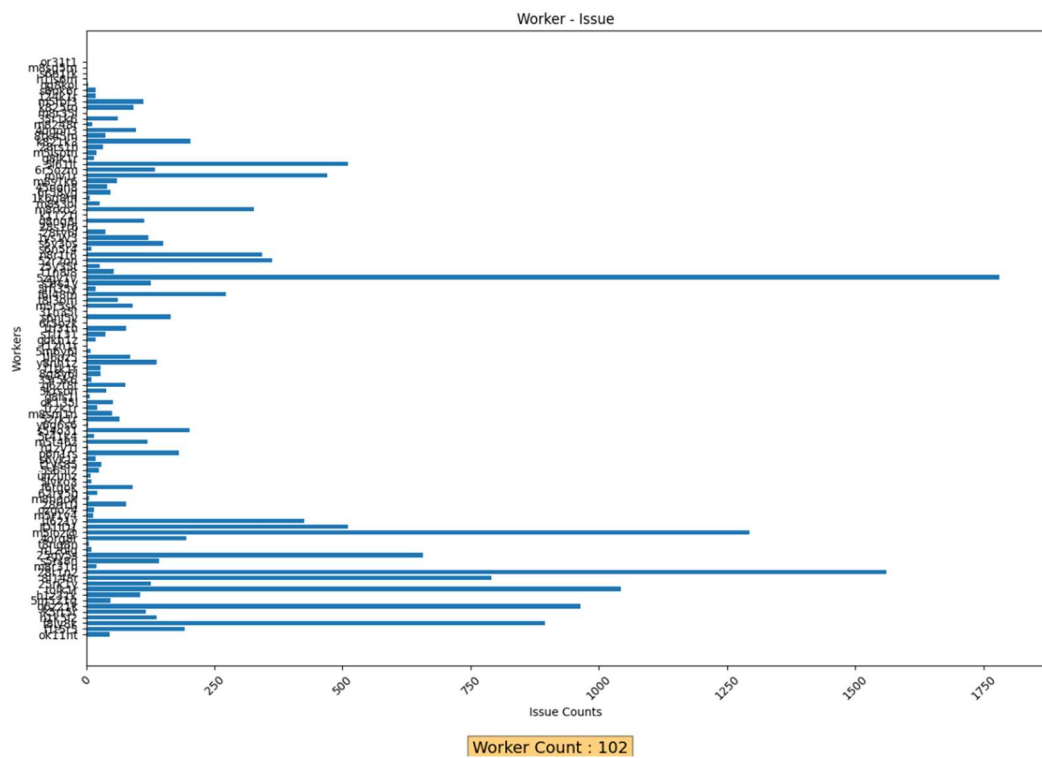
- Employee types and their counts:
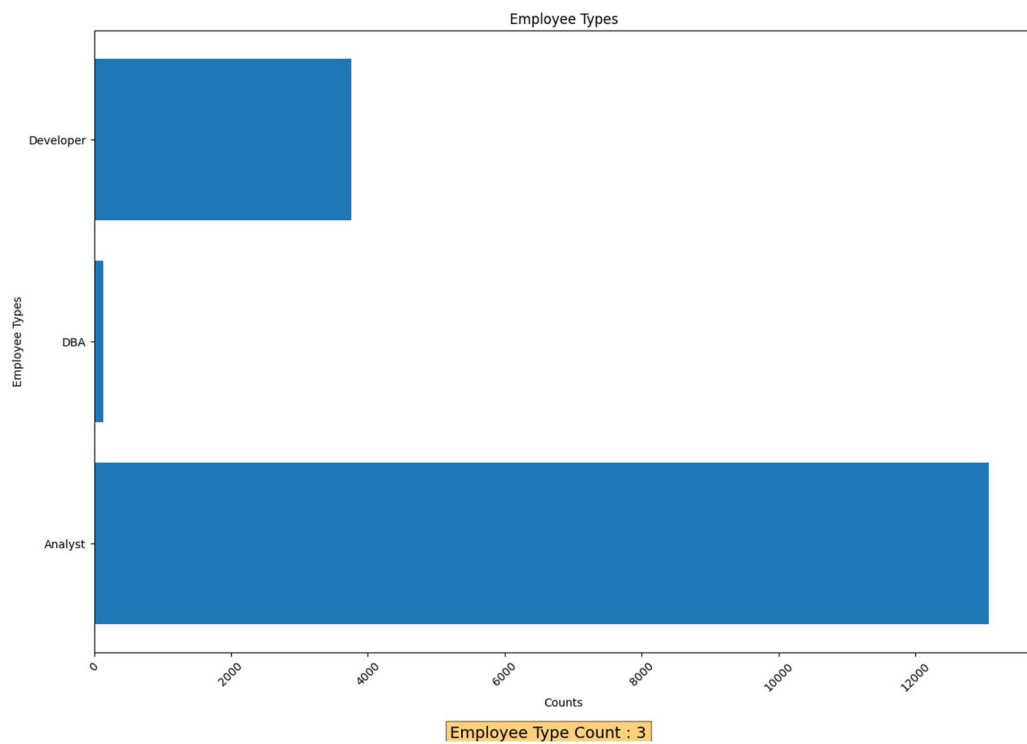
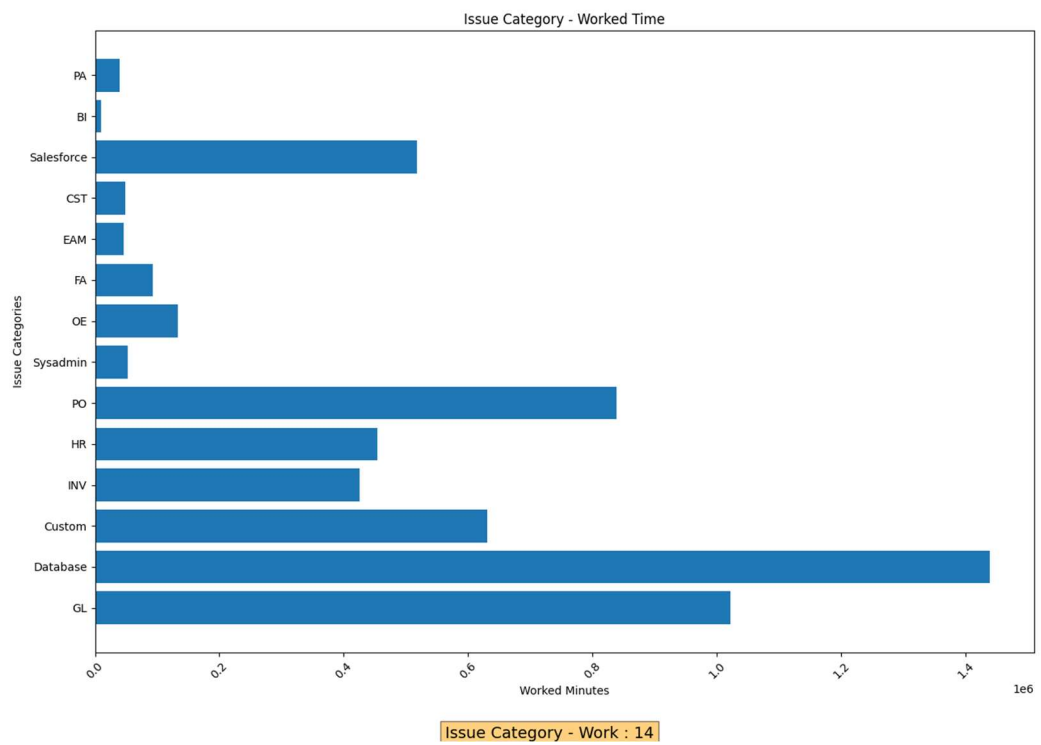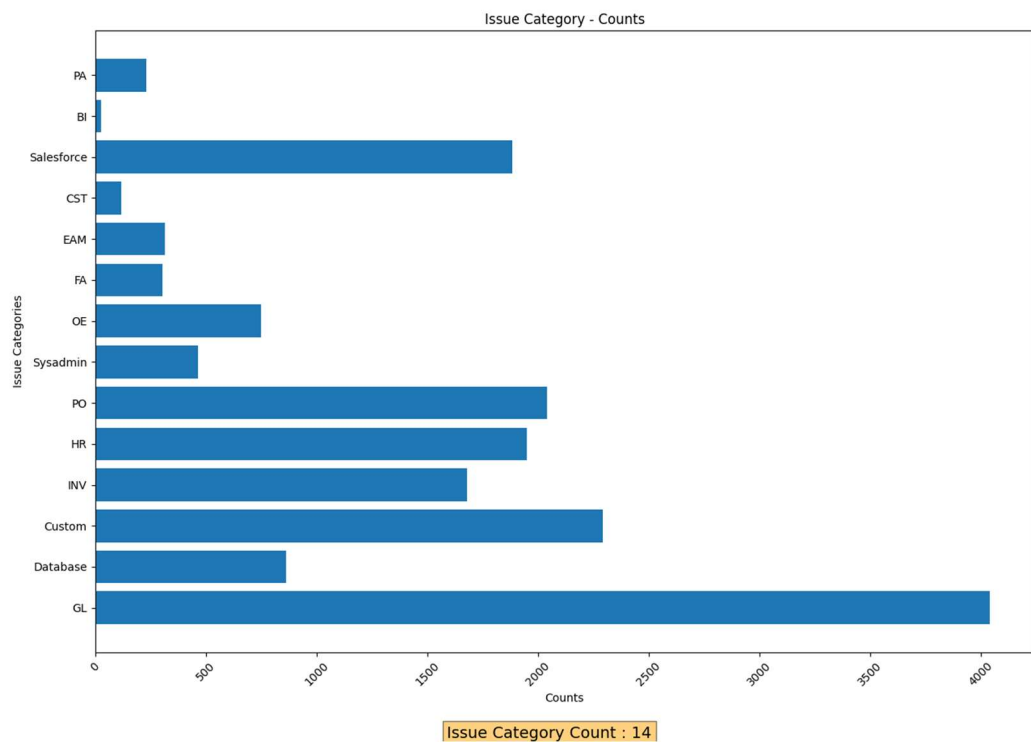- Number of issues that each employee worked on:

- Total minutes spent on each issue type:

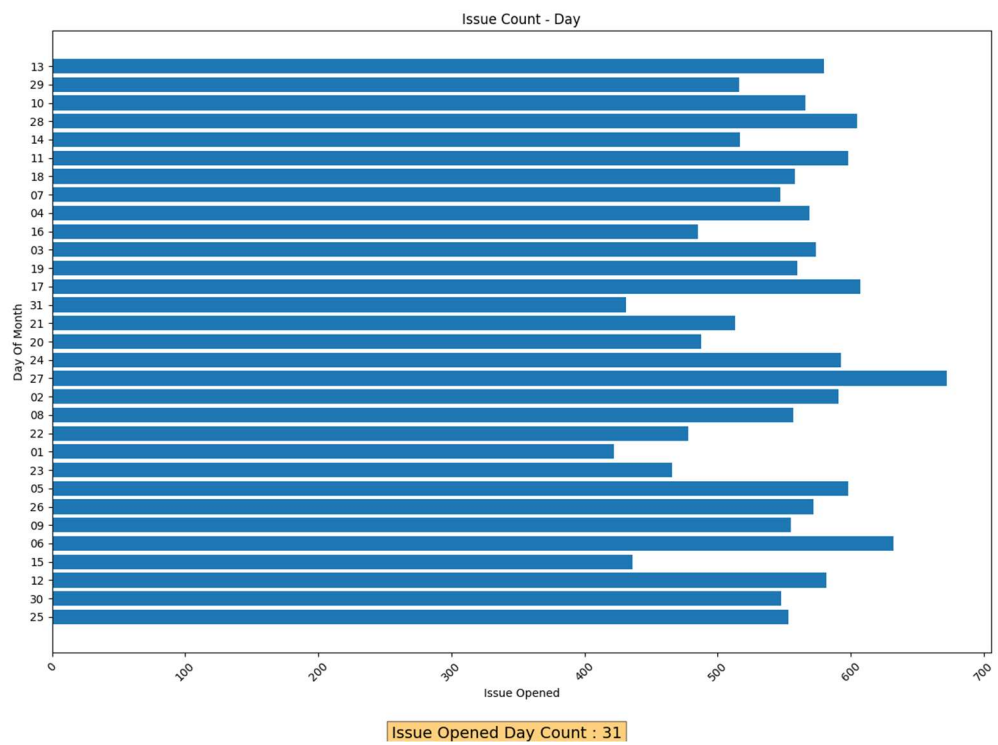- Total minutes spent on each issue category:

Issue Category - Work : 14

- Number of each issue category:

Issue Category Count : 14

- Number of issues opened at days of a month:

Issue Opened Day Count : 31

- Number of issues opened at months of a year:

Issue Opened Month Count : 12