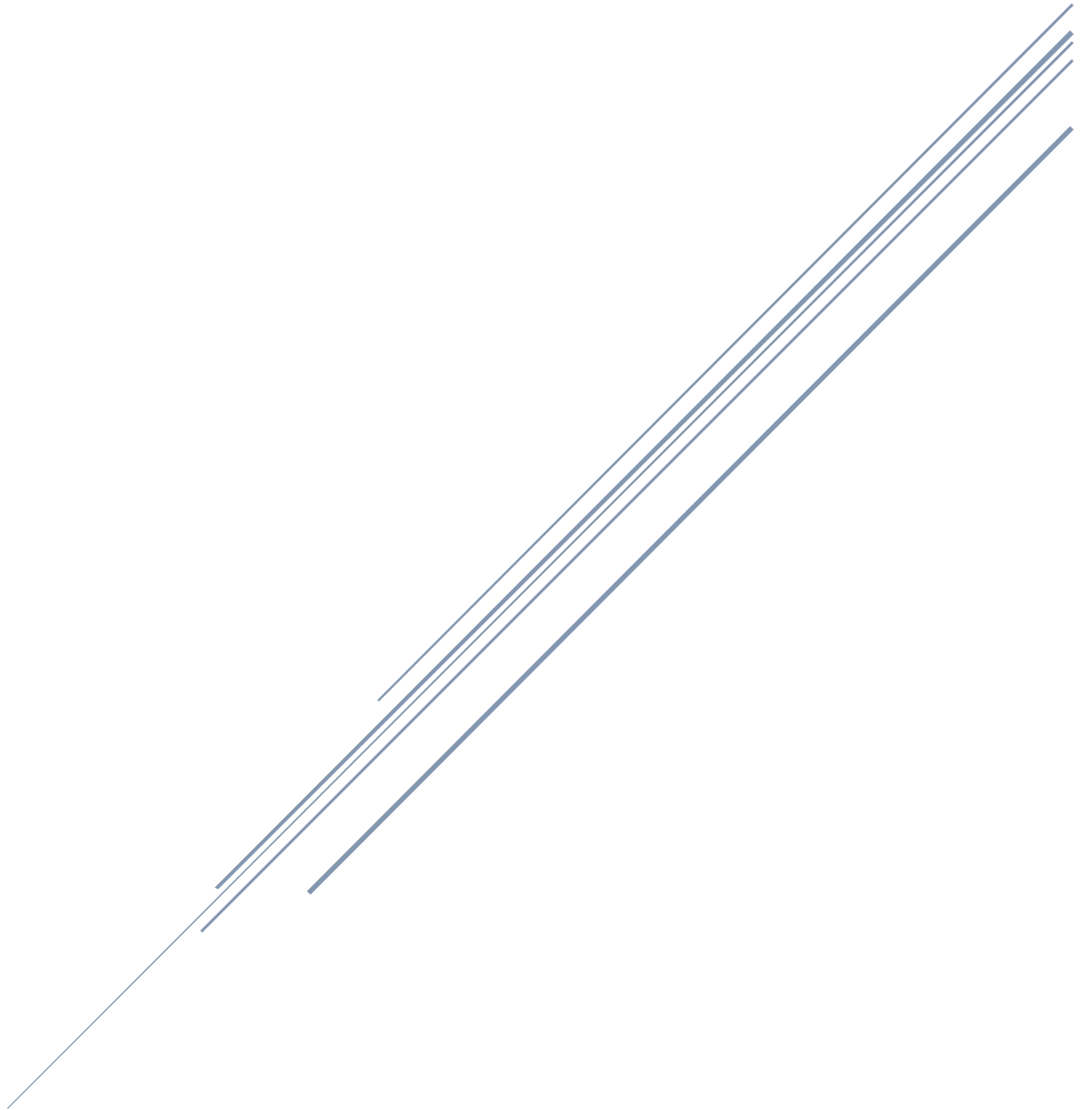# LINEAR-REGRESSION PROJECT PROPOSAL

SDAIA T5 Data scince boot camp

Norah Alqahtani, Albatoul Alosaimi, and Shroaq Almutairi

Introduction:
This is the second project for the T5 Data Science Bootcamp, which is about building a linear regression model that address a useful prediction using Scikit-learn, and using data scraped from website with Requests, BeautifulSoup, and Selenium. This project helps ReadMe firm which is an online web site that is selling books. In order to determine the price of its products they ask for our help as data scientists to study the books features to predict a suitable price.

Question/need:
1. What is the suitable price for books given available features?

Data Description:
we aim to produce a model that is able to investigate book features such as title, authors, average_rating, isbn, isbn13, language_code, num_pages, ratings_count, text_reviews_count, publication_date, and publisher. In order to determine a suitable price.
We collect the data from two deferent data sources, first source is a CSV dataset publicly available at: https://www.kaggle.com/jealousleopard/goodreadsbooks it has originally more than 11,000 rows and deferent 12 columns. We select 1298 rows and 12 columns.
The second source is the price of these books we gain it from https://www.bookfinder.com we use web scrap selenium to fetch this information. Taking advantage of selenium feature that allow the browser to interact with the web site specially in our case we make selenium searching for book prices by its isbn. Then we save the results into csv file and merge it with our previous dataset.

Algorithms
- We build a linear regression in python, and techniques such as regularization and polynomial features, adding interaction terms and dummy variables have been used.
- Rigorous model selection and evaluation has been used to select model between Linear Regression, Ridge Regression, Polynomial Regression, and Lasso Regression

Tools
- selenium in python (Scraping the web)
- Pandas and Numpy (Exploring the data)
- Matplotlib and Seaborn (Visualizing the data)
- Sci-kit Learn (linear Regression model and other models)