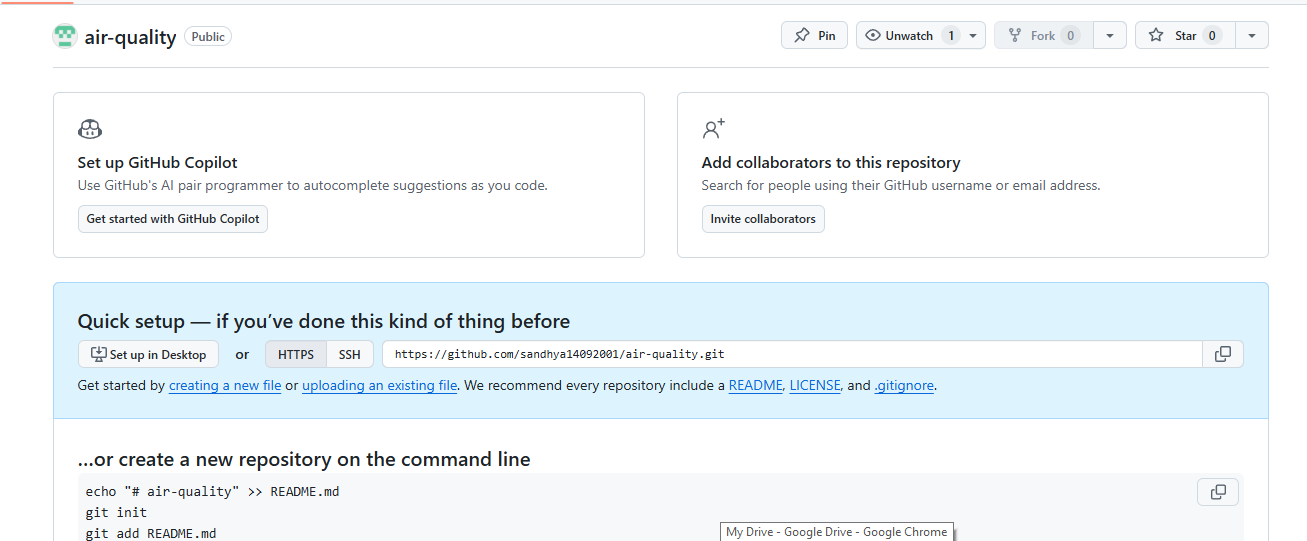
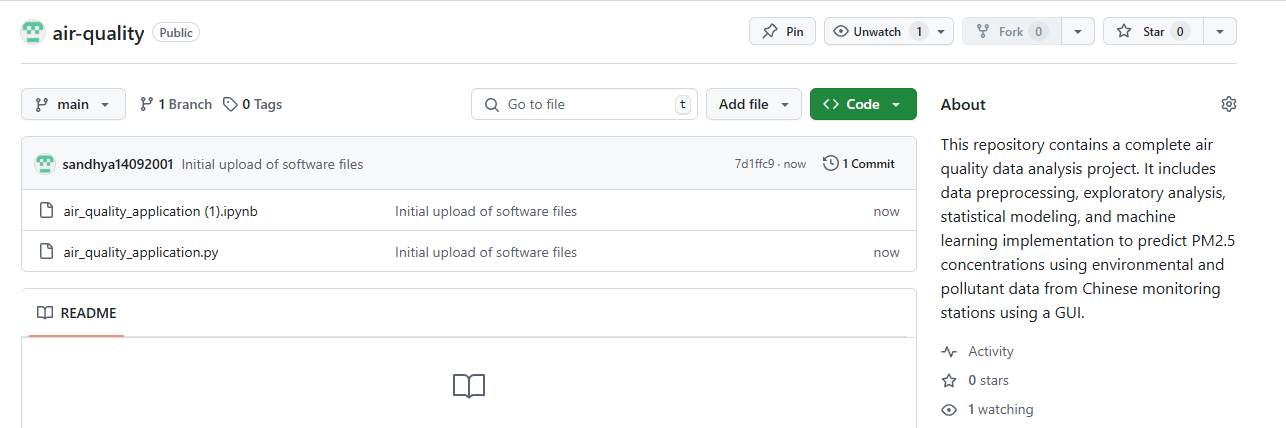
# Version Control



**Figure 29: Github repository**

(Source: Self-created)



**Figure 30: Adding files into Github repository**

(Source: Self-created)

Version control has been executed through creating a GitHub repository and regularly committing the project’s progress. The first commit included creating a basic layout of the project and loading the dataset, data cleaning and exploratory data analysis (EDA). Further commits showed the inclusion of linear regression and the random forest model, as well as the incorporation of the evaluation parameters. Further, commits were made on improving the UI where the Gradio platform was integrated to allow users to peek at summaries and trends and also make predictions. Further commits documented enhancements in the model quality, say, the hyperparameter adjustments. The benefit of committing is that it provides a record of the progression of the project within the documentation trails, which can easily be accessed from the changes made from time to time.

# Discussion

The results show a systematic approach to analysing the data collected from different stations concerning air quality in China. The dataset was accumulated effectively and combined into one structure based on hourly values of pollutant and meteorology parameters (Nath *et al.* 2021). In order to get a better understanding of the dataset, various EDA procedures were conducted on the dataset, including missing values, statistical distribution, and the relation between pollutants and the weather conditions.

In the preprocessing stage, null values have been handled and removed, as well as the construction of some additional feature variables, including a datetime variable to support time series analysis. This was succeeded by simple distributions and simple associations that pointed out trends of, for instance, PM2.5 variation with months and the inverse proportional relationship between temperature and pollutants. This was also evident in the heatmap and pairplots, as they clearly showed interconnections of pollutants and how pollution is dispersed depending on the wind speed and direction of different sources (Kothandaraman *et al.* 2022). Model construction included building linear regression and random forest regression models with inputs of environmental factors to predict PM2.5 levels. The linear regression model produced an R² of 0.84, suggesting a good fit, and the random forest increases the effectiveness of the classification. The study conducted overall gives crucial information concerning the general trend of air quality and the effects of weather factors on pollutants.