# Summary of Data Analytics Projects in Python

# Sean Zhang

## In this summary, only pure python projects are mentioned. Data preprocessing and Data Manipulation for other tools such as Tableau, Power BI, and Excel, are not included.

## Exploratory Data Analysis (EDA) & Data Visualization Projects

1. **Small Projects:**

|  |  |  |
| --- | --- | --- |
| **Project Name** | **Project Notes** | **Tools** |
| [Data Science related positions Analysis](https://www.kaggle.com/seanzhang1997/ba-da-or-ds-plotly-tells-you) | 1. data cleaning 2. data visualization (interactive plot) | Pandas, NumPy, Plotly, Matplotlib, Seaborn. |
| [Data Science Jobs visualization and word cloud](https://www.kaggle.com/seanzhang1997/plotly-interactive-visual-define-a-good-job) | 1. data cleaning 2. data visualization (interactive plot) | Pandas, NumPy, Plotly, Matplotlib, Seaborn, NLTK. |
| [Airbnb Dataset](https://github.com/zhangzixuan1997/Python/blob/master/small%20projects%20_%20data/Airbnb%20Dataset%20Analytics.ipynb) | 1. Data Merging and Cleaning with Python PandasSQL. 2. data visualization (interactive plot)  3. Prescriptive Data Analytics | Pandas, NumPy, PandasSQL, Matplotlib, Seaborn |
| [Windows Store EDA and Visualization](https://www.kaggle.com/seanzhang1997/windows-store-eda-and-visualization) | EAD and Visualization | Pandas, Matplotlib, Seaborn |

## Machine Learning & Deep Learning Projects

1. **Big Projects:**
2. [**Humana Health Care Case Competition**](https://github.com/zhangzixuan1997/Python/tree/master/Humana%20Case%20Competition)**.** Manipulated professional health care domain large datasets (more than 800 features). Applied data cleaning, visualization, feature engineering (feature transformation, feature selection with Random Forest, clustering with T-SNE and PCA, data sampling technique), machine learning modeling (Logistic Regression, Random Forest, SVM, XGBoost, Ensemble Method, Sequential Neural Network). Drafted data analytics report.
3. [**Kaggle Competition**](https://github.com/zhangzixuan1997/Python/tree/master/Kaggle%20Competition%20House%20of%20Blocks) **–** Houses of Blocks Falling Prediction. *I can only share the basic code because the competition is still on going*. Applied Convolutional Neural Network (CNN) model on over 14GB and 50k images to predict block structures' stability. Built generator for image pre-processing (brightness and sharpening) and applied Transfer learning with EfficientNet, ResNet101, and Xception. The model reached 0.002 loss with hyperparameter tuning. The model can be extended to practical object stability assessment.
4. [**Veganism Industry Research and Text Classification**](https://github.com/zhangzixuan1997/Python/tree/master/Veganism)**.** Dealt with a large dataset containing vegetarian restaurants data across the country. Performed data cleaning, feature engineering and data visualization. Performed text cleaning and text classification on menu text.
5. **Small Projects:**

|  |  |  |
| --- | --- | --- |
| **Project Name** | **Project Notes** | **Tools** |
| [Topic Modelling with LDA detecting Medium Articles’ Topics](https://www.kaggle.com/seanzhang1997/topic-modeling-lda-which-topic-has-more-claps) | 1. EAD and Visualization 2. Text Cleaning & Topic Modeling (LDA Analysis). | Pandas, NumPy, NLTK, LDA. |
| [Fake News Detection](https://www.kaggle.com/seanzhang1997/99-3-data-endogenous-led-to-result-word2vec-nn) | 1. EDA, Visualization. 2. Text Cleaning  3. Text Data Feature Engineering with Tokenization and Vectorization(w2v). 4. Simple Neural Network | Pandas, NumPy, NLTK, Word2Vec, Keras. |
| [Hotel Review Text Mining](https://github.com/zhangzixuan1997/Python/blob/master/hotel%20review%20text%20mining.ipynb) | 1. Text Data Cleaning 2. Text data tokenization and lemmatization. 3. Lexicon Unsupervised Learning 4. Feature Engineering with TF-IDF 5. SVM and Random Forest for text mining 6. Topic Modeling 7. Confusion Matrix study and threshold analysis. | Sklearn, NLTK, Pandas, NumPy |
| [Health Care – Heart Attack Possibility](https://www.kaggle.com/seanzhang1997/logistic-svc-rfc-neural-network-nn-is-the-best) | 1. EDA, Visualization. 2. ML with Logistic Regression, SVM, Random Forest and Neural Network. | Pandas, NumPy, Matplotlib, Sklearn. |
| [Student Performance](https://github.com/zhangzixuan1997/Python/blob/master/Student_Performace-A%20simple%20deep%20learning%20early%20stopping%20example.ipynb) | Neural Network Sequential Model with KFold. | SciPy, TensorFlow, Keras, NumPy, Pandas. |
| [MINIST Dataset](https://github.com/zhangzixuan1997/Python/blob/master/MINIST_CNN.ipynb) | Neural Network CNN Model with Classic image classification problem. | TensorFlow, Keras. |
| [Housing Dataset Multiclassification with Deep Learning](https://github.com/zhangzixuan1997/Python/blob/master/Multiclassification.ipynb) | Neural Network Sequential Model with Housing Dataset. AUC and ROC for measurement | TensorFlow,  Matplotlib,  Keras. |
| [Google Cloud Platform Python ML](https://github.com/zhangzixuan1997/Python/tree/master/GCP) | GCP course project with TensorFlow and Keras. | TensorFlow,  Keras,  GCP. |
| [H2O AutoML](https://seanzhang19971026.medium.com/h2o-automl-in-python-comprehensive-tutorial-f25001c11b80) | AutoML with H2O. Apply multi-ML models on classic Titanic dataset. | AutoML, H2O. |

## Python Programming and Prescriptive Analytics Projects

1. **Big Projects**
2. **Customer Data Crawler and Prescriptive Analytics with SciPy package.** For work reason, the code cannot be shared. During my work at 3Trees, I scraped customer data with BeautifulSoup, urllib, re, xlwt and MySQL; launched data analytics program over store-level data and applied SCAN-PRO model with regression. I utilized Python SciPy and Pandas packages for ROMI sensitivity analysis and operational research to maximum cost reduction.
3. **Python data manipulation and Prescriptive Analytics.** For work reason, the code cannot be shared. During my work at BCG, I was responsible for importing, exporting, manipulating data (data cleaning and basic visualization) with Python. Further Excel data were generated by Python for consulting purposes.
4. [**Python Scraping Movie Data.**](https://github.com/zhangzixuan1997/Crawler_with_Python) Scraped data from Douban with BeautifulSoup, urllib, re, xlwt and sqlite3. Procedure: build get data functions with RegEx, get the URL by manipulating log in information for anti-crawler, and build save data function to export the data to SQLite database and Excel.
5. [**Small Projects**](https://github.com/zhangzixuan1997/Python/tree/master/Prescriptive%20Analytics)

## Other Projects related with Data Analytics

* 1. **Medium Author**
     1. [*Vegetarian/Vegan Restaurants Data Analysis with Python*](https://seanzhang19971026.medium.com/vegetarian-vegan-restaurants-data-analysis-with-python-bc12623f657a)
     2. [*AdaBoost Algorithm*](https://seanzhang19971026.medium.com/boosting-technology-for-machine-learning-adaboost-add021b997bc)
     3. [*Gradient Boosting Algorithm*](https://seanzhang19971026.medium.com/boosting-techniques-for-machine-learning-gradient-boosting-for-regression-and-classification-283c5d13853f)
     4. [*XGBoost Algorithm*](https://seanzhang19971026.medium.com/boosting-techniques-for-machine-learning-xgboost-for-regression-and-classification-507376eedd6f)
     5. [*XGBoost Optimization and Hyperparameter Tuning*](https://seanzhang19971026.medium.com/boosting-techniques-for-machine-learning-xgboost-optimization-and-hyperparameter-tuning-38b18861d02)
     6. [*Practical Applications of ML Models*](https://seanzhang19971026.medium.com/practical-applications-of-machine-learning-models-when-and-where-to-use-which-model-6477dfb1c384)
     7. [*H2O AutoML in Python Comprehensive Tutorial*](https://seanzhang19971026.medium.com/h2o-automl-in-python-comprehensive-tutorial-f25001c11b80)
     8. [*Advanced SQL–Function, Procedure, Trigger, CTE and Recursive CTE*](https://seanzhang19971026.medium.com/advanced-sql-222e32a7fa37)*.*
  2. **[LeetCode](https://leetcode.com/zhangsean1997/)** –Database Questions 78/136 completed (on going).