# Pengchong Tang

Kalamazoo MI 49009 Email: ferrarisf50@gmail.com Phone: (269) 365-6057 Github Linkedin

## SUMMARY

4+ years of solid working experience on SAS programming and Data Analysis. Substantial hands-on experience with SQL, Python and R. Master of Science in Statistics and Bachelor of Science in Math.

## SKILLS

| | |
|---|---|
| **Programming languages:** | SAS, R, Python, SQL, MATLAB, Scala, VBA, C# |
| **Data science & big data Tools:** | Hadoop, Pig/Hive, Spark, PySpark, MongoDB, Tableau, AWS, BigQuery, R Shiny |
| **Version control systems:** | Git, SourceTree |
| **Databases & platforms:** | Oracle, SQL Server, Linux |
| **Technology knowledge:** | Machine learning, Data Analysis, Data Mining, Data Wrangling, Predictive Modeling, Statistics, Natural Language Processing (NLP), Operations Research, Experimental Design, MapReduce, HDFS |

## EXPERIENCE

**Eurofins Lancaster Laboratories /** SAS programmer/Scientist II at Kalamazoo, MI **11/2014 – present**

- Support clients with animal health data management, design SAS/Python/SQL scripts to build data mappings, transcribe data from multiple sources (Internal, EDC, lab result) and various formats (Excel, XML, SAS, SQL Server, flat file) into SAS tables, manipulate, merge, concatenate, transpose and clean SAS tables using PROCs and DATA steps, automate 60% of the mapping pipelines by implementing machine learning algorithms (tf-idf) to classify lab spreadsheet data. (Main skills: SAS/MACRO, SAS/BASE, SAS/SQL, Scikit-learn, NLP, Random forest, SQL Server)
- Define data validation rules, revamp the old procedure by designing SAS and Python programs to detect the data issue systematically during the mappings, investigate data issues, perform ad hoc data analysis and report findings to the clients. (Main skills: PROC SQL, PROC SUMMERY, PROC FREQ, DATA STEP, etc.)
- Design and develop a Python dashboard, which utilizes packages such as Selenium, PyODBC, and Win32com.client to automate repeated data management tasks, including email management, classifying mass PDF files and web forms filling, successfully reducing labor hours for clients by 80%. (Main skills: VBA, Python, Web scraping)

## PROJECTS

**Credit Card Fraud Detection** **11/2017**

- Explored and analyzed 280 thousand credit card transactions using R ggplot2 and reported findings using R Markdown.
- Identified transaction patterns and created univariate, bivariate and multivariate EDA reports.
- Built a classification model using Scikit-learn in Python. The model achieves 0.8 score on average on AUPRC, can detect about 80% of frauds.

**NYC Taxi Trip Duration** **01/2018 - present**

- Wrangled 1.4 million NYC taxi trip routes information during the first half of 2016 from OSRM, assessed and cleaned the data using Pandas/Numpy, created data visualization dashboard with Tableau, Matplotlib and Seaborn, performed data analysis using Spark on AWS EMR.
- Queried more than 100 million NYC taxi trips using BigQuery, visualize the findings.
- Implemented machine learning algorithms (Xgboost) using Scikit-learn to build a model to predict taxi trip duration, scoring LB 0.38.

## EDUCATION

- MS in Statistics, George Washington University, Washington DC **09/2010 – 06/2012**
- BS in Mathematics, Sun Yat-sen University, Guangzhou, China **08/2006 – 06/2010**

## CERTIFICATIONS

Udacity Data Analyst Nanodegree Program, SAS Certified Base & Advanced Programmer for SAS 9, R Programming (Coursera), Data Visualization and Communication with Tableau (Coursera), Hadoop Platform and Application Framework (Coursera) (Completed)