An

Internship Report

Name of the student: Aditya Kaushal

UID: 16BCS7098

at

Aerogram Pvt Ltd, Synergy Building, IIT Delhi, Campus, Hauz Khas, New Delhi, Delhi 110016

Submitted in Fulfillment of

Bachelor of Engineering Computer Science

(Cloud Computing)

in

Apex Institute of Technology

Chandigarh University, Gharuan

(Batch No. 2016-20)

Submitted by: Aditya Kaushal.	**Guided by: Manoj Sahukar**
Student Name: Aditya Kaushal	**Reporting Manager: Manoj Sahukar**
UID: 16BCS7098	**Designation: Data Science/Engineer Intern**
BE CSE (Cloud Computing)	**Semester VIII**

Executive Summary:

About Company:

Aerogram is an Indian IIT Delhi Incubated start-up that is devising a network to predict real-time air quality in a local mapped area. Aerogram uses its own built sensors that measures Air Pollution metrics such as PM 2.5, PM 10, Temperature, Humidity and pressure. Around more than 50 sensors are placed in the vicinity of IIT Delhi, Hauz Khas. Aerogram is building citywide network of air pollution monitors to track personal exposure to pollution. The Aerogram is led by **Dr. Sarita Alahwat (CEO)**. Some of their devices named are **Ezio Sense, Ezio STAT, Ezio Motiv**.

Objective:

The main Objective of the Internship was to Prototype a Flask Public Web-App Dashboard for Time Series Analysis of Particulate Matter 2.5 for Forecasting and Prediction of PM 2.5. The aim of this was to get future values of Particulate Matter 2.5, so to generate quality insights on a click of a button. Furthermore, to understand the MQTT (Message Queuing Telemetry Transport) Architecture for the sending the telemetry feed from the IoT devices (Sensors) collecting PM 2.5, Temperature, Pressure, PM 10, and Humidity to Google Cloud IoT Core and Pub/Sub for storing the data into Google Cloud Firestore and Google Cloud SQL Instance. The Job also consisted to understand the various fluctuations of PM 2.5 during, and before the national lockdown due to the Covid-19. The internship helped me to gain exposure to Machine Learning, Python Programming Language and it's frameworks like Flask.

Methodology/Tools Deployed:

The utilization of NumPy, Pandas, Matplotlib, Seaborn, Cufflinks, Time Series Forecasting Algorithms like (ARIMA, SARIMA and FB Prophet (Facebook's Prophet)), Statistical Components, Tableau, Google Cloud Platform, Google IoT Core, Google Cloud Functions, Google Pub/Sub and Data Analytics during the Internship period helped me to gain various practical exposure to all these tools and frameworks.

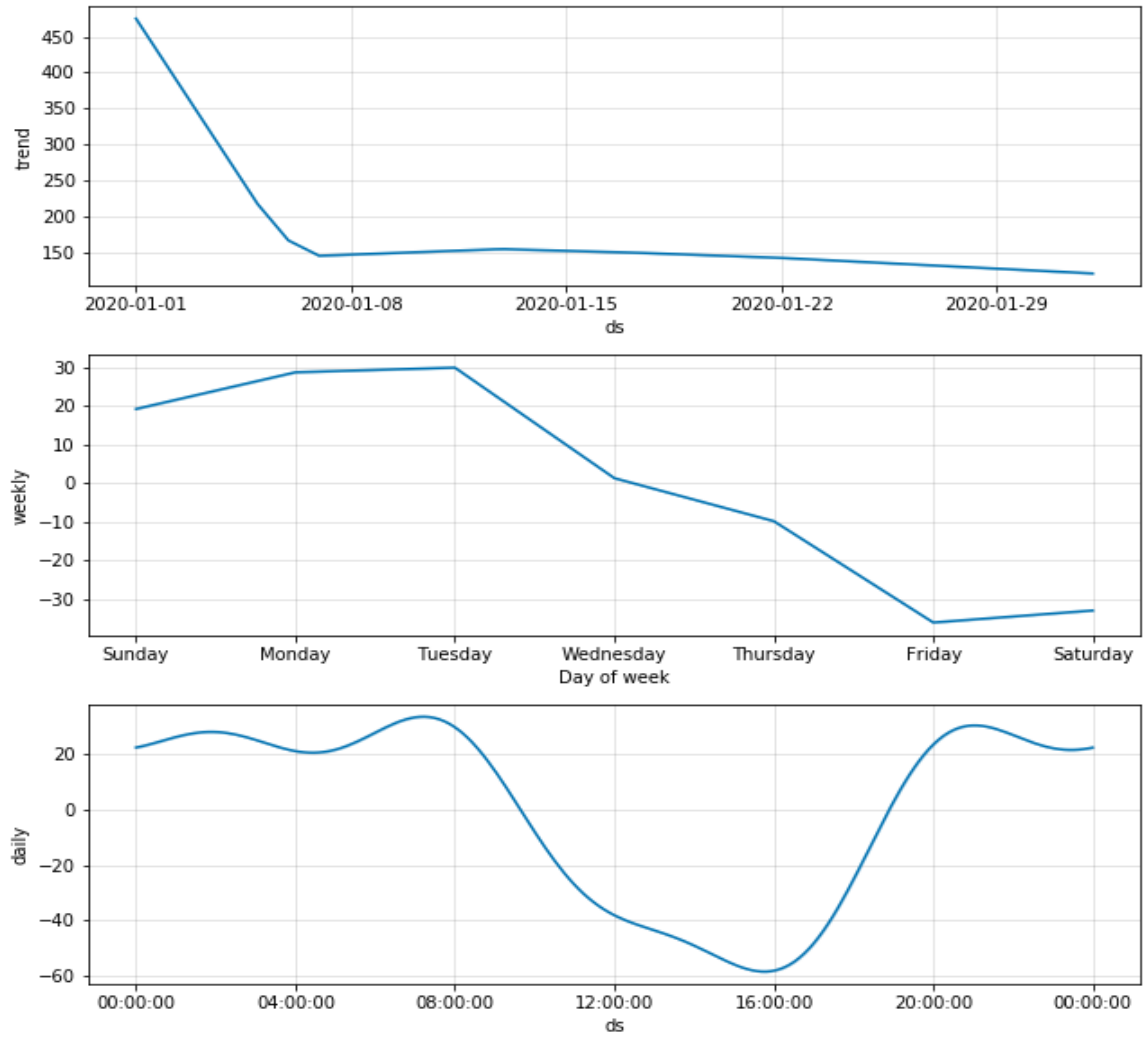
Key parts of the report (Findings and solutions):

- Time Series Analysis for Forecasting and Predicting the future PM 2.5 Values.
- Investigation of various individual sensor location metrics such as PM 2.5, PM 10, Temperature and their correlation with the BAM Data.
- Understanding and doing extensive research on fluctuation of PM 2.5 Values during and before the lockdown due to covid-19.
- Understanding and determining the seasonality and trend of PM 2.5 values during the day and night hours of a day.
- Deploying a Flask web app which integrates a web-site dashboard. The dashboard also has the various components like daily averages of PM 2.5, lowest mean, highest mean. The web dashboard mainly comprised of the feature for forecasting and predicting the PM 2.5.

Benefits to the company / institution through report.

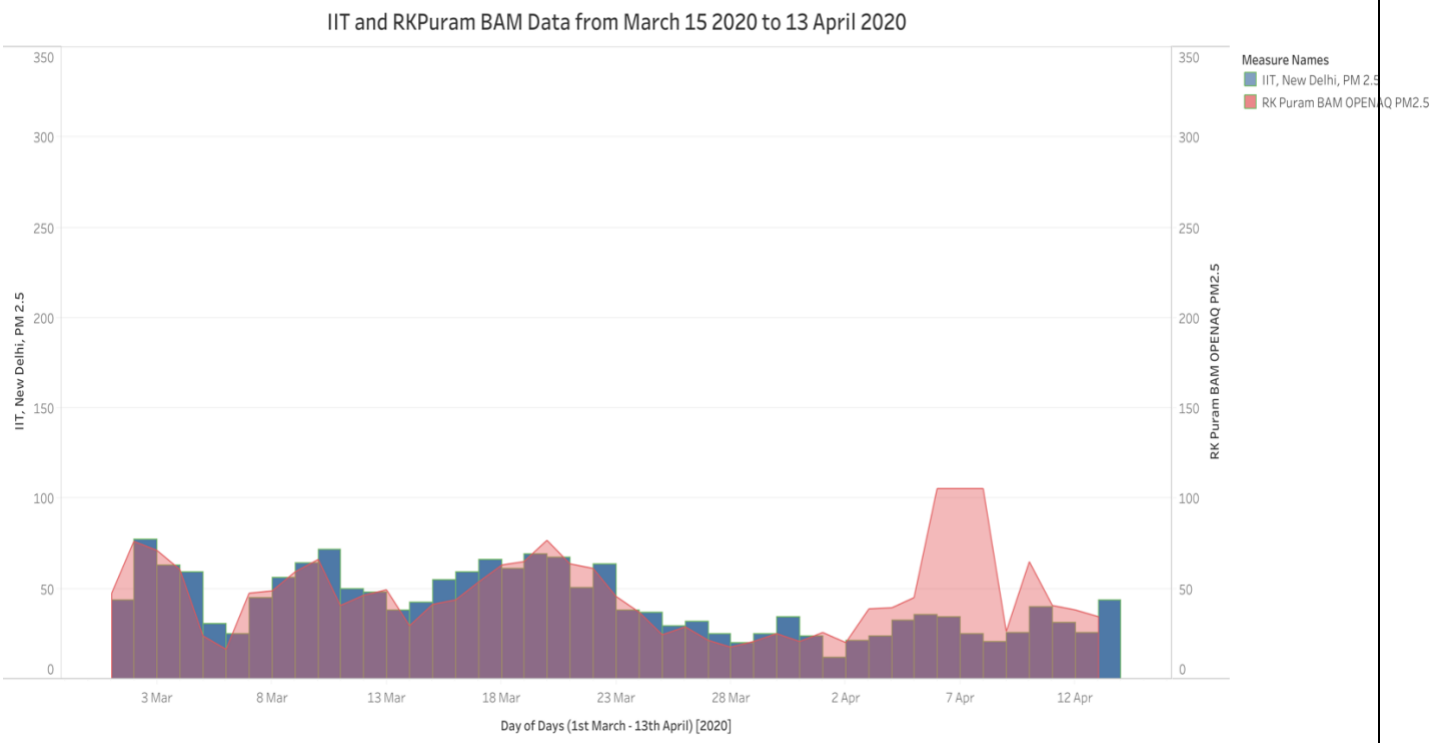
- The work carried out during the Internship period in Aerogram Pvt Ltd, was very fruitful and full of learning about practical exposure to different technologies such as Data Analytics, Data Engineering, Cloud Computing, Python Programming, Machine Learning, Statistics, generating various insights and understanding about Particulate Matter 2.5 and its harmful effects to humans.
- The job that was carried out in Aerogram was to find out the correlation between PM 2.5 and various other Air Pollution metrics such as Temperature, Pressure, Humidity, PM 10.
- Investigation about fluctuations of PM 2.5 values in the IIT Delhi Campus with respect to BAM (Beta Attenuation Monitoring) put by the CPCB (Central Pollution Control Board) in various locations of Delhi like RK Puram, IGI Airport, and different locations in East South, Central and North Delhi, but not limiting to these.
- Carried out work on Time Series Forecasting using various algorithms like ARIMA, SARIMA, FB Prophet for making predictions on PM 2.5 values 3-8 hours ahead for generating insights about how the PM 2.5 exposure would affect the health in the upcoming hours of a particular day.

All this work has been very beneficial to Aerogram and many IIT Delhi professors already working on the Air Pollution effects. The work carried out in Aerogram was also complied into a report consisting of Plots and Graphs representing how the PM 2.5 values and other air pollution contributors fluctuate during the various hours of the day and night, so as to know which part of the day contributed the most to PM 2.5. It also consists of reports and plots on trends and seasonality depicting how the PM 2.5 changes during the different hours of the Day.

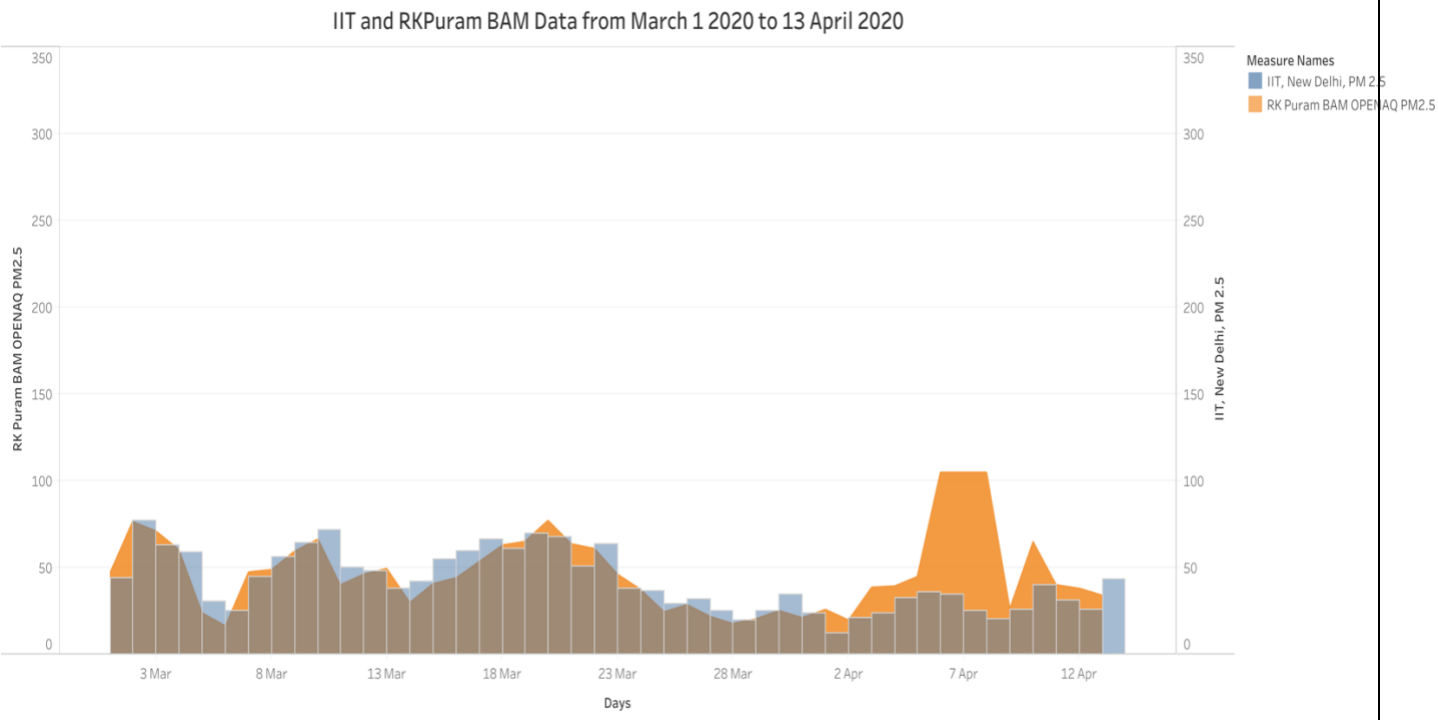


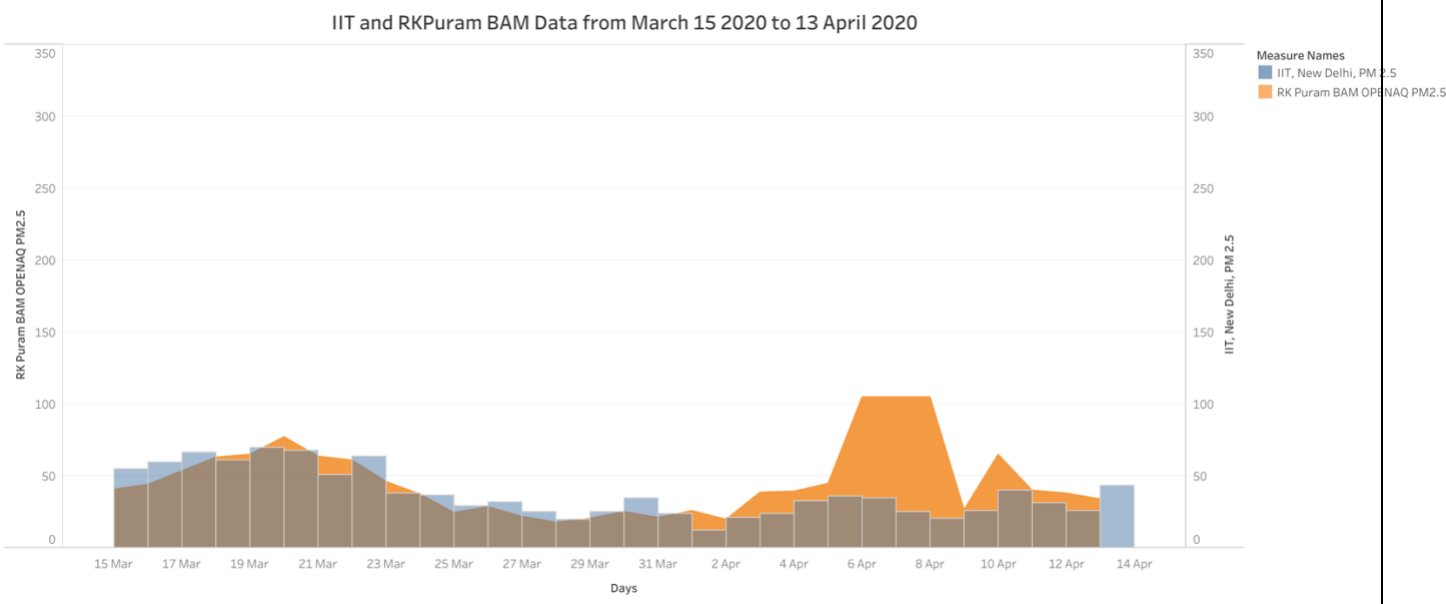
This is a Seasonality, Monthly Trend and Weekly Trend of how the PM 2.5 Value changes during the different hours, week and months' time.

The report furthermore consisted of plots and graphs which depicts the comparison of how the PM 2.5 values varied during the national lockdown and before the national lockdown due to the unprecedented covid-19 situation. The plots cover the month of January, February, March, April.



This is the plot of Sensors in IIT put by Aerogram compared with the RK Puram BAM data.



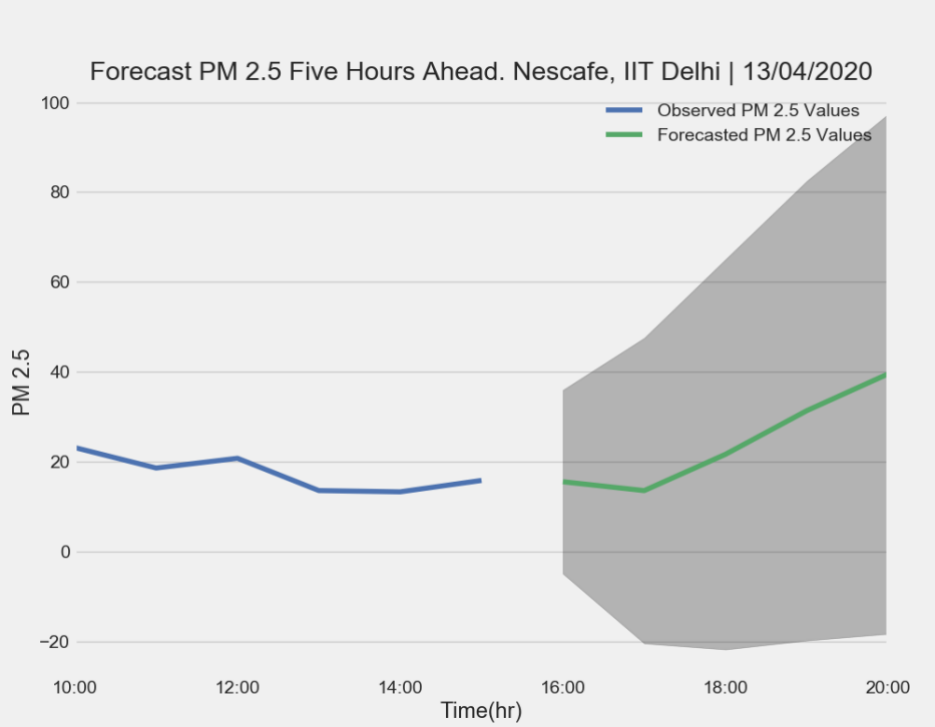


This is the plot for the month of April (Lockdown)

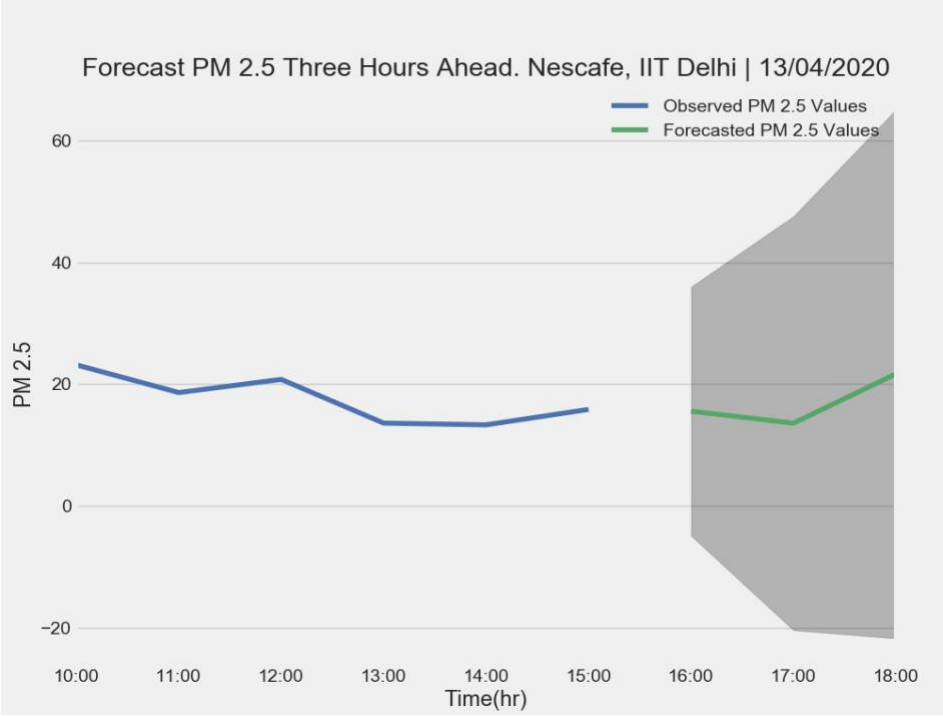
The work done during the Internship has been beneficial to me and to Aerogram. The work carried out has been able to find out many insights as following:

- The correlation between the Particulate Matter 2.5 and Temperature during the Day and Night.
- Information about the seasonality, trend of Particulate Matter 2.5.
- Insights on how the PM 2.5 fluctuated during and before the national lockdown.
- Developing of a Flask Web-App for generating a web dashboard for the consumption by public/consumers to forecast and predict various data points of PM 2.5 and also plotting the last 6 hours actual data and the predicted/forecasted 3 hours data.
- Analyzed the various tools like Google Cloud Platform, Google IoT Core, Google Pub/Sub to study about various methodologies to send and receive telemetry feeds.
- The various data that has been generated by the sensors was also cleaned and feature engineered.
- The utilization of the various tools like Tableau helped to plot quality graphs for understanding the various factors of Particulate Matter 2.5.
- The various libraries of Python, Pandas, Matplotlib, Cufflinks helped to understand the raw data and get insights from it.
- The Forecasting algorithms like ARIMA, SARIMA and FB Prophets are based on many statistical concepts. Using these statistical concepts, the Forecasting and prediction of PM 2.5 was done.

The below graph is the Forecasted and Actual Value plots for a location in the IIT Delhi Campus. This is the Plot for Nescafe in IIT Delhi for 3 hours and 5 hours ahead.



3 Hours ahead.



5 Hours ahead.

Table of Contents / Index Page:

Introduction to Company and Industry:

Aerogram is focused on establishing a network of low cost high quality air pollution monitoring sensors which helps individuals know what they breath on real-time-basis. Aerogram's infrastructure is based on data collection, analysis and distribution wherein they gather historic data, alert on real-time data and conduct analysis to predict future readings of identified air pollutants in our environment.

Products and Services:

Aerogram has its products and services in the area of solving and managing the problem of Air Pollution and especially doing extensive research on the Particulate Matter 2.5. Aerogram has three products at the current which has been made available to the public for consumption. The products made are utilized for capturing the Air Pollution metrics such as: Temperature, Pressure, PM 2.5 and PM 10. The products namely are: Ezio-Motiv, Ezio-Sense, Ezio-Stat.

1. The **Ezio-Motiv** is a device which is mounted on the Bus, and it is used for measuring metric such as PM 2.5, and temperature of the location wherever the bus goes and stops for passengers.
2. The **Ezio-Stat** is a device which are placed in the vicinity of the IIT Delhi Campus.
3. The **Ezio-Sense** is a personal home device which can be used for measuring the Particulate Matter 2.5 and PM 10.

Products and Services of Aerogram

These are Ezio-Sense, Ezio-Stat and Ezio-Motiv.

Latest Developments:

Aerogram has its foot in many government funded projects for measuring the contribution of PM 2.5 by stubble farming. Aerogram is working on new devices which will be used to measure the contribution of Smoke and Particulate Matter 2.5 through the neighboring states.

Other project also involves the use of Ezio-Motiv devices mentioned above, which are mounted on the Delhi Buses for collecting the PM 2.5 emissions around the city.

Data Collection:

Aerogram also collects Data fetched using the 40-50 Ezio-Stat Devices represented using the multi-colored circular nodes in the image given below. These devices are made by Aerogram (an IITD Incubated Startup) which gathers and fetches various measurements like Temperature, Pressure, PM 2.5, PM 10 and etc. These devices are used to micro-map the entire Indian Institute of Technology - Delhi Campus to collect and use the data for further use such as Time Series Analysis, prediction, and forecasting. The Scale varies from 50<= PM 2.5 <=250, and the various colors represent and change these values according to the given scale. Each and every Ezio-Stat Sensor is placed 100mts apart. The micro-mapping shows that at every 100mts the PM 2.5 values can vary.\\\

Detailed Introduction to your job profile:

Designation: Data Science/Engineer Intern

Job Location: IIT-Delhi, Synergy Building, New Delhi, Hauz Khas.

Job Description: As a Data Engineer Intern, the specific job role was to be able to design, build, operationalize, secure, and monitor data processing systems with a particular emphasis on security and compliance. The main objective was to work on Time Series Analysis for forecasting and predicting the Particulate Matter 2.5 values in real time. Using Statistical Tools and algorithm to understand the implementation of algorithms using Python/R and deploy a web-dashboard for the public consumption. Furthermore, the job was related to Data Analytics for getting insights about the fluctuation of PM2.5 values during the lockdown and before the lockdown.

Key Performance Areas:

- Utilized the Time Series Forecasting to predict and forecast PM 2.5 values and other Air pollution metrics using regression analysis.
- Developed Python scripts for publishing and subscribing to telemetry data to utilize the MQTT Architecture to send and receive Air Pollution metrics on Google Pub/Sub and Google IoT core.
- Analysed the real-time sensor data generated by the Ezio-Stat devices (developed by Aerogram used for micro mapping the IIT-Delhi campus) for fetching the PM 2.5 values and analysing the seasonality, trend, and the noise component to further understand the various factors that contribute to the fluctuations of PM 2.5 in correlation with temperature, pressure, and humidity.
- Prototyping a public Web dashboard by using HTML, CSS, JS, Flask, and Cloud Firestore for the end-user for forecasting PM 2.5 value a few hours ahead so as to provide insights.
- Developed Google Cloud Functions using Python as a programming language to migrate the telemetry feed received from sensors to Google Cloud Pub/Sub, Google Cloud SQL, and Google Cloud Firestore.
- Determined the weekly and monthly comparison of the Air Quality Metrics such as PM 2.5 during and before the COVID-19 lockdown.
- Utilized Tableau to plot the monthly and weekly comparisons of BAM (**Beta attenuation monitoring** (**BAM**) is a widely used air monitoring technique employing the absorption of beta radiation by solid particles extracted from air flow) and the Aerogram Sensors to crosscheck values with respect to Central, West, North, and East Delhi.
- Determined the Correlation between the Aerogram Sensors and BAM. Utilized the Scatter Plots and Heatmaps for determining the visualization representing the Correlation.

- Performed feature engineering and data cleansing from raw data for performing analysis on the metrics provided by the Aerogram Sensors.
- Resampled the Raw Data into Hourly, Weekly, and Daily Interval Samples for analysing and determining the Day-wise Maximum and Minimum.
- Investigated the trends, seasonality of the PM 2.5 for the Month of December 2019, (January, February, March, April) 2020 for understanding fluctuation of the Pollution Metrics in correlation with effects of national lockdown starting from March till present.
- Utilized the FB Prophet Algorithm to understand the Seasonality and trend of the PM 2.5 during the Day and the Night.
- Carried out Analytics for plotting the graphs of PM 2.5 values with respect to the BAM (Beta Attenuation Monitoring) Data taken from OPEN AQ for RK Puram and Aerogram's individual sensors placed in the vicinity of IIT Delhi Campus (e.g. Nescafe Cafeteria, LHC (Lecture Hall Complex), Seminar Hall, Block 6 (3rd Floor), Kusuma School, Wind T).

Internship Timeline/ Daily Work Hours:

The Internship timeline was from 15th January, 2020 to 15th May, 2020. The Daily office work hours were from 10:00 AM to 5:30 PM.

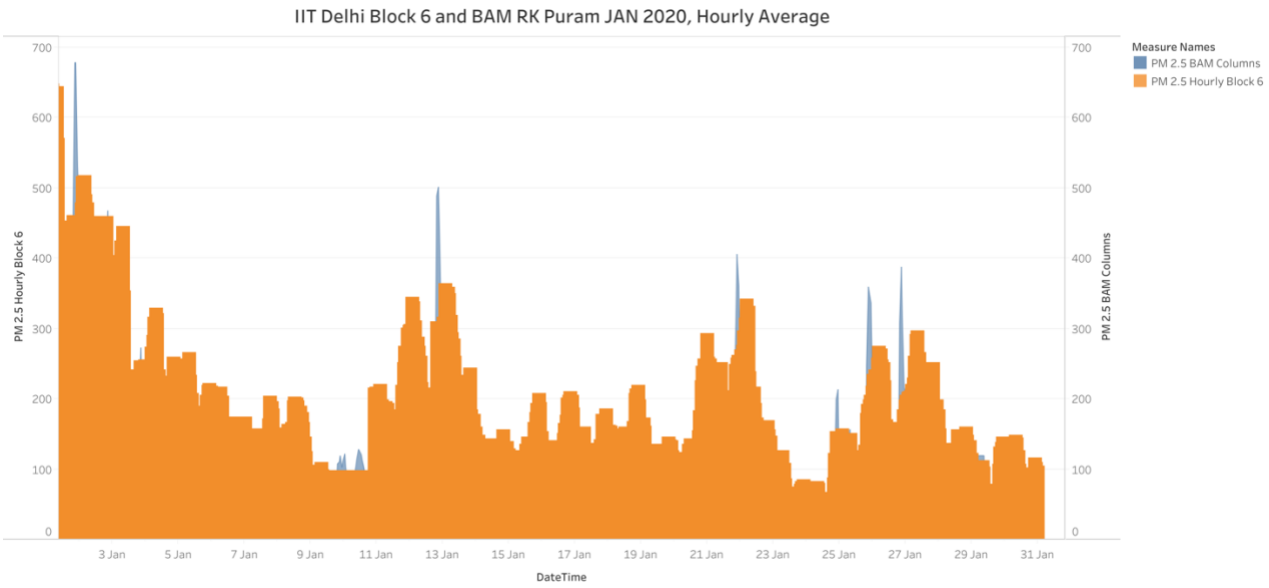
Detailed Job Description:

As a Data Science/Engineer Intern, the specific job role was to be able to design, build, operationalize, secure, and monitor data processing systems with a particular emphasis on security and compliance. The main objective was to work on Time Series Analysis for forecasting and predicting the Particulate Matter 2.5 values in real time. Using Statistical Tools and algorithm to understand the implementation of algorithms using Python/R and deploy a web-dashboard for the public consumption. Furthermore, the job was related to Data Analytics for getting insights about the fluctuation of PM2.5 values during and before the lockdown. Furthermore, the role also demanded to be able to be well equipped with Cloud Computing for the utilization of Google Cloud Platform. Also, the understanding of various tools like Tableau for plotting quality and interactive plots. Day-to-Day task also demanded to be able to implement various forecasting algorithms to develop Forecasts for Particulate Matter 2.5.

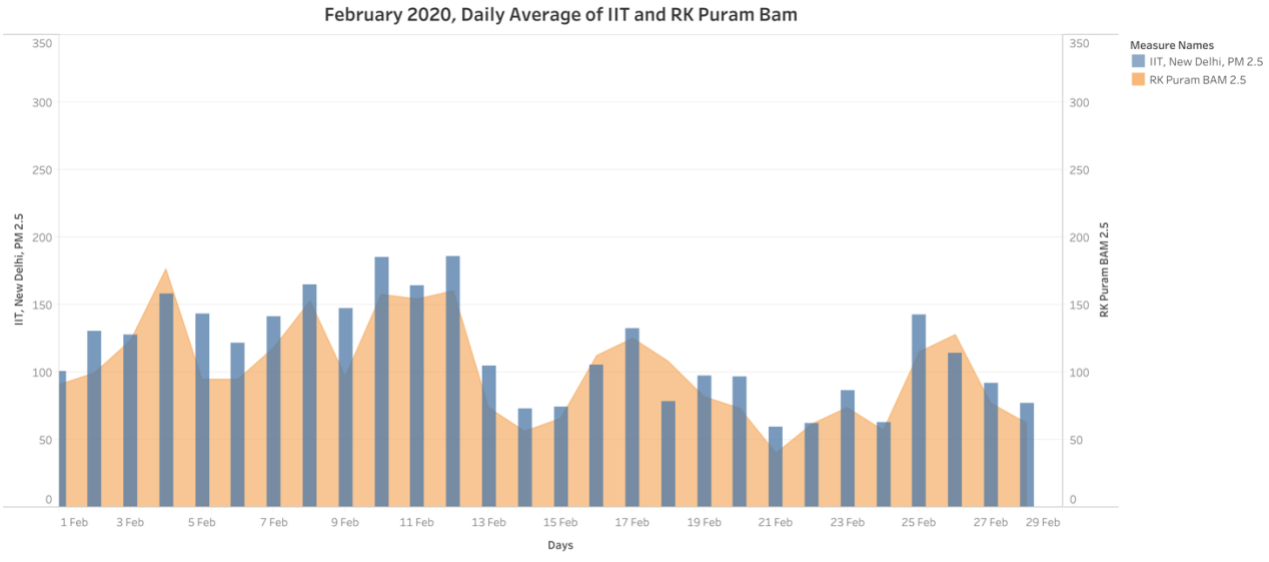
Daily Task:

- Implemented various forecasting algorithms to develop 3 Hourly Forecasts for Particulate Matter 2.5.
- Build a Web-Dashboard to be able to deploy the algorithms to a web application using Flask, HTML and CSS.

- Utilization of Tableau Public to build quality and interactive charts for understanding the various fluctuations in PM 2.5. The Hourly, Weekly, Monthly comparisons were done for the months of Dec 2019, Jan 2020, Feb 2020, March 2020, April 2020.

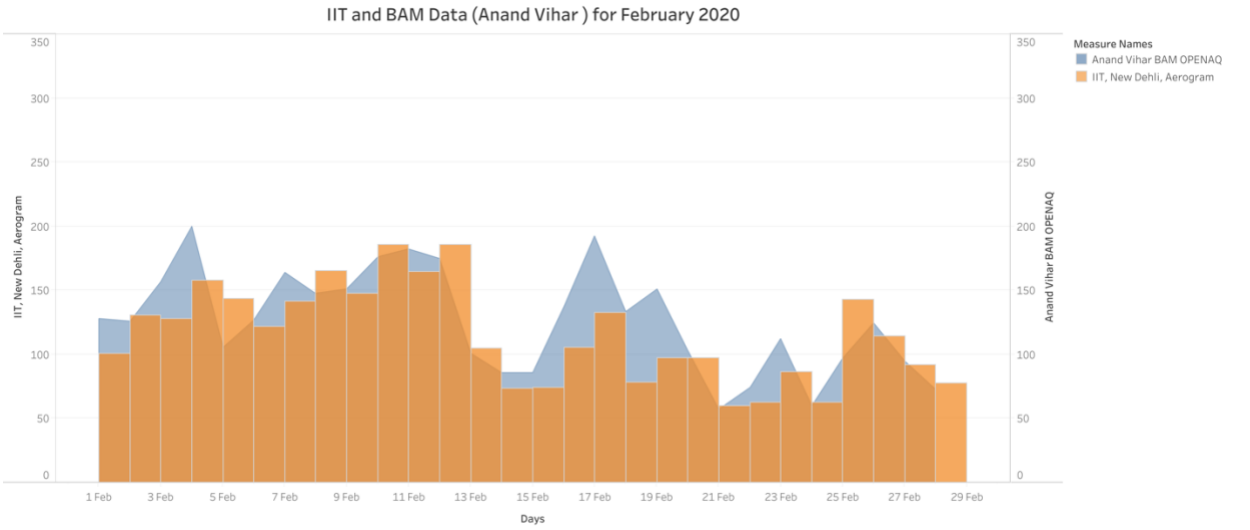


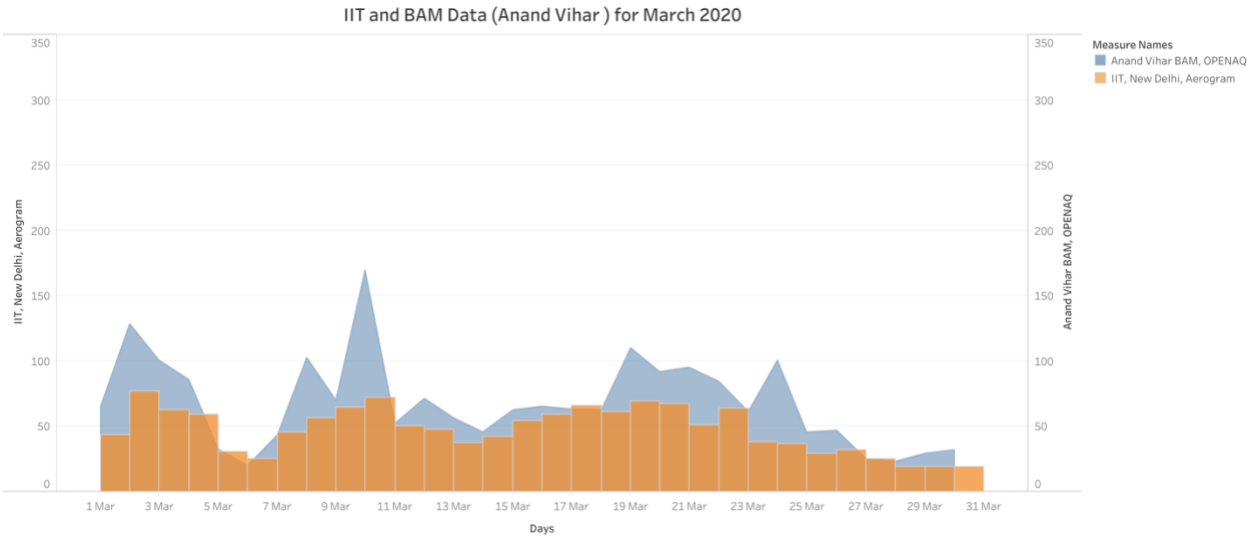
Hourly Plots for Block 6 IIT Delhi Campus in comparison with the RK Puram Bam.



Daily Average of IIT and RK Puram

- Also, compared the overall Average of all the sensors Particulate Matter 2.5 values in the vicinity of IIT Delhi with the RK Puram BAM (Beta Attenuation Monitoring), BAM in central, west, east, south and north Delhi.





- Investigated and analyzed various patterns and alteration in the Daily Minimum and Maximum values of the PM 2.5 values created from the Raw Data.

DateTime	pm25Max	pm25Min
2020-03-01	69.96581196581197	17.128205128205128
2020-03-02	98.7008547008547	49.61864406779661
2020-03-03	90.99145299145299	22.713043478260868
2020-03-04	107.38260869565218	18.23008849557522
2020-03-05	39.042735042735046	6.205128205128205
2020-03-06	39.758620689655174	8.628318584070797
2020-03-07	69.11965811965813	16.99137931034483
2020-03-08	73.27966101694915	24.025641025641026
2020-03-09	167.59322033898306	15.81578947368421
2020-03-10	189.83760683760684	27.771186440677965
2020-03-11	52.47457627118644	35.939655172413794
2020-03-12	122.2457627118644	11.273504273504274
2020-03-13	63.3728813559322	12.626086956521739
2020-03-14	52.0	23.796610169491526

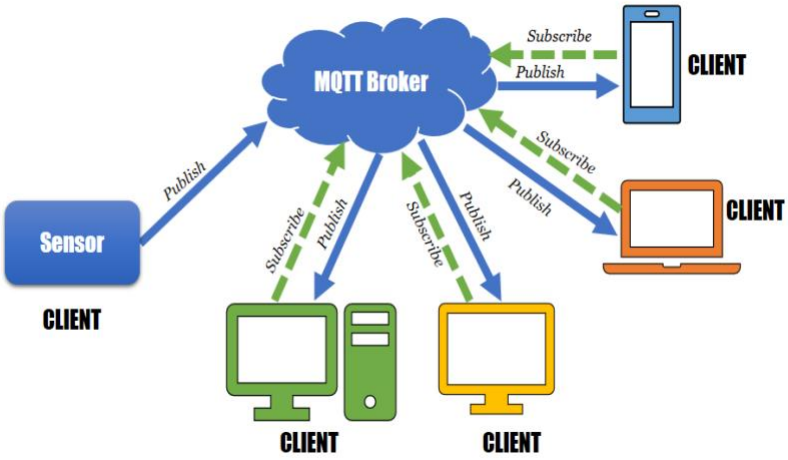
DateTime	pm25Max	pm25Min
2020-03-01	79.53846153846153	23.305084745762713
2020-03-02	105.76271186440678	47.00854700854701
2020-03-03	111.03389830508475	17.042735042735043
2020-03-04	101.96610169491525	13.82905982905983
2020-03-05	43.495726495726494	5.818965517241379
2020-03-06	48.085470085470085	9.405172413793103
2020-03-07	93.38135593220339	19.068376068376068
2020-03-08	92.80508474576271	17.29059829059829
2020-03-09	140.9322033898305	10.085470085470085
2020-03-10	194.87394957983193	26.440677966101696
2020-03-11	66.7542372881356	37.059322033898304
2020-03-12	117.82203389830508	10.76271186440678
2020-03-13	59.17094017094017	9.586206896551724

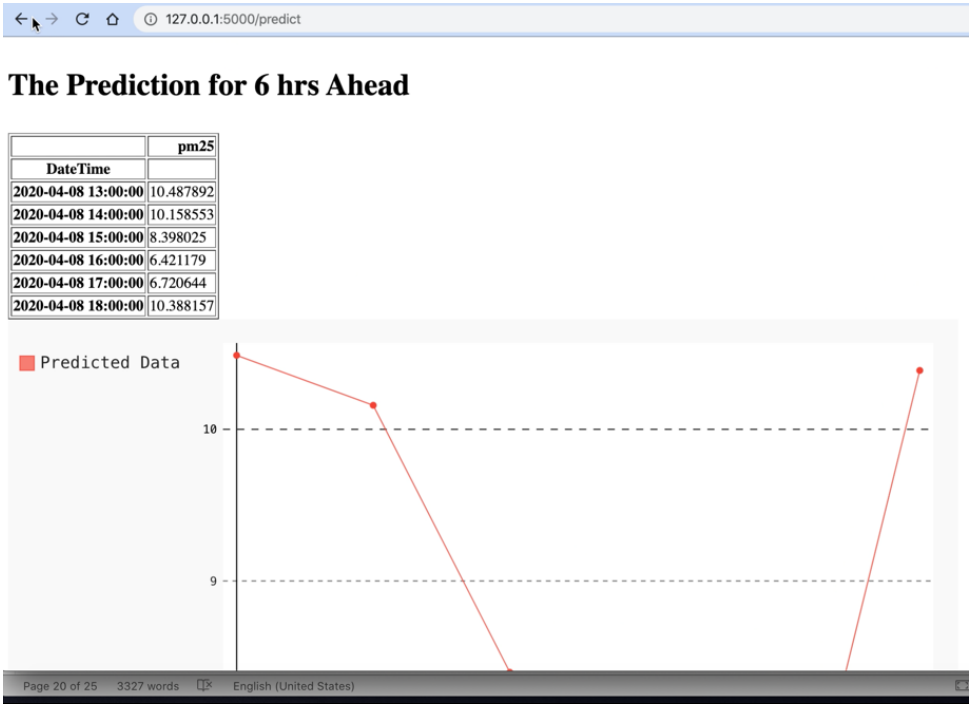
Key Learning from Internship: The major learning and take-a-ways during the duration of the Internships were as following:

- Better Understanding of Data Analytics, Python Programming Language and Machine Learning. The work done in Aerogram was extensive and full of research. This led to better grasping of concepts in Python Programming Language and its various libraries used for mathematical computing, graph plotting, statistical plotting and data handling. The various libraries like Pandas, NumPy, Matplotlib, Seaborn, cufflinks were used for the above-mentioned tasks
- Determined and studied about various Time Series Forecasting algorithms like Moving Averages, Exponential Smoothing, Holt-Winters Method, ARIMA (Auto Regressive Integrated Moving Averages), SARIMA (Seasonal Auto Regressive Integrated Moving Averages) and FB Prophet for predicting the Particulate Matter 2.5.
- Utilized the Tableau Public for Data Analytics and inbuilt forecasting for building quality and interactive plots for comparing monthly, weekly and daily averages of Particulate Matter 2.5.
- The implementation of Python framework Flask, was also undertaken to implement and deploy a web-dashboard.
- Deployment of MQTT Architecture utilized for publishing and subscribing the telemetry feeds from the IoT Devices to MQTT Broker.
- Understood the various commands of version controlling for committing my work in different git repository.

The architecture of MQTT (Message Queueing Telemetry Transport)

From communication between devices and collection of data the protocol used was MQTT (Message Queuing Telemetry Transport). The MQTT is used for low-bandwidth IoT Devices. The MQTT uses the concept of Publish and Subscription, Topics and an MQTT Broker. The MQTT Broker acts as a central server for subscribing and publishing to particular topics. The Topics are the location where the messages would be published by the IoT sensors and subscribed to the devices which require the published data. The telemetry data would only be published to the devices which are subscribed to the topic. This type of architecture utilizes the bandwidth in an efficient way. The MQTT utilizes the TCP\IP Protocol for the connection between the devices and the MQTT Broker.





Internship/ Project Discussion:

How the objectives were achieved?

The objectives were achieved using Python Programming language, Tableau as tool for Data Analytics, Statistical Concepts such as Time Series Forecasting, Mean, Median, Mode, Variance, Averages, cross validation and regression analysis but not limited to these. Furthermore, utilization of Google Cloud Platform services such as Google IoT core, Google Cloud Pub/Sub were also made for implementing an MQTT Broker for sending and receiving the telemetry feed from and through sensors collecting air pollution metrics such as Temperature, Humidity, Pressure, PM 2.5, PM 10.

What skills (scientific and professional) were learned during the internship?

The day-to-day task involved working on Python Programming Language and implementing the various algorithms for Time Series Forecasting. Using Python as Programming I wrote Google Cloud functions to automate tasks and create a workflow on Cloud. The utilization of NumPy, Pandas, Matplotlib, Seaborn, cufflinks also led to the addition of skills. Investigating the fluctuations and altering of Particulate Matter 2.5 values also led me to use Tableau as a tool. Further skills like, Data Analytics, using Google Cloud Platform, Git Commands, Cloud Services, statistical concepts like Moving Averages, ARIMA, SARIMA also led to gain of skills.

Results / observations/work experiences get in the internship company.

PM2.5 particles are fine particles with a diameter of 2.5 micrometers or smaller. They occur mainly through combustion (such as stubble burning, vehicular emissions). PM2.5 particles can travel deeper into your respiratory system and even enter your bloodstream, leading to coughing, tightness in your chest and wheezing. In some cases, it can also lead to severe diseases such as chronic obstructive pulmonary disease (COPD) and cancer. They are the primary polluters in India. Exposure to fine particles such as Particulate Matter 2.5 ($PM_{2.5}$) can travel deeply into the respiratory tract, reaching the lungs, which can cause short-term health effects. The main objective is to carry out a Time Series Analysis and create a public web-app dashboard that will forecast the Particulate Matters 2.5 in real-time using the Time Series Analysis done on observed data (past data). Further, the dashboard would provide the user interface to the public to predict and analyze the forecasts of the Particulate Matter 2.5, **'X'** number of hours ahead on a click of a button. The Benefit of such a project will be to create awareness among the public, providing real-time data access to the public, gain insights and important notifications about the severity of the Air Pollution in their neighboring area. It would also suggest avoiding places with high PM 2.5 values leading to the safety of the public. The Web-App would also include Charts and Plots for the end-user interaction and for the ease of providing the information to the end-user.

What challenges did you experience during the internship?

1. Implementing the Time Series Forecasting Algorithms
    a. The implementation of the algorithms like SARIMA, ARIMA, Moving Averages and finding the algorithm which correctly predicts the PM 2.5 Values was challenging.
    b. Data cleaning and feature engineering the data was also challenging.
2. Creating combination charts consisting of Line plots and Bar Plots together involved a lot of research.
3. Understanding how to implement the security in MQTT through OPENSSL certificates by using Public and Private Key.
4. Data Cleaning and selecting potential features for Time Series Forecasting.
5. Feature Engineering in the raw data.
6. Understanding the various concepts of Statistics and Probability.
7. Understanding the Python Paho MQTT Client and its uses.

Conclusion:

Getting an internship in Aerogram Pvt Ltd led me to gain a practical exposure in field of Data Analytics. This opportunity also led me to understand what kind of challenges are faced while working in a startup. This journey made me more interested and excited to get involved into a real professional arena. This experience brought out my strength and also the areas I needed to make up. It added more confidence to my professional approach built a stronger positive attitude and taught me how to work in a Team as a player. The primary objective of an internship is to gather a real-life working experience and put their theoretical knowledge into practice. This was my first real life experience to work on Air Pollution Analytics and extensively work on Cloud, Data Science, Forecasting and predication algorithms and much more.

I also learned the values and importance of the industry and experienced that this much superior field than most of the other field during my training. As a human being, I noticed many changes in my attitude. I am more confident and more likely to do any work now.

During my Internship, I thoroughly enjoyed the challenges that came along every single day. I learned that this is the just the beginning of the road and I have to travel a long distance to be a successful person in this field. But I must say that this experience will prove an objective in my career in the IT Industry.

References:

- https://medium.com/@serbelga/build-a-weather-station-with-google-cloud-iot-cloud-firestore-mongoose-os-android-jetpack-350556d7a
- https://medium.com/@gguuss/google-cloud-iot-core-to-cloud-sql-8d194a28fc7f
- https://pypi.org/project/google-cloud-pubsub/
- http://www.steves-internet-guide.com/into-mqtt-python-client/
- https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
- https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06
- https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/
- https://towardsdatascience.com/serving-prophet-model-with-flask-predicting-future-1896986da05f
- https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-prophet-python-r/