Zuha Ahmad

DS-670: Capstone: Big Data & Business Analytics

Dr. Jaume

04 February 2017

**ASSIGNMENT 5**

**Method (Algorithm Overview and Implementation in Zeppelin)**

**OBJECTIVE**

The objective of the project is to execute a statistical modeling by performing a regression analysis and a time series forecasting of pollution based on traffic sensors. We will use programming tools and techniques, specifically the Zepellin notebook, to import the geo-spatial and environmental data, inputting a superlative algorithm, and processing the data with visualization. The contributing variables in this dataset are ozone index levels, carbon monoxide, nitrogen dioxide, sulfur dioxide, and particulate matter. By observing the trend and seasonality associated with each pollutant, we can determine which chemicals causes more harm, check the low and high peaks of the season, and come up with business ideas to reduce cost.

**DATA COLLECTION**

The pollution dataset derives from two locations. The first is from Aarhus, Denmark and the second is Brasov, Romania. The timeline of the dataset is taken in the year 2014 for a few months for each country. The raw dataset is in the CSV which has 81,012,652 bytes. However, the annotated dataset is distributed in the gzipped file format which has 3,368,454,165 bytes. The structure of the dataset is structured because it has numerical fields and timestamps.

The contributing variables in this dataset are:

* Ozone index levels
* Carbon monoxide
* Nitrogen dioxide
* Sulfur dioxide
* Particulate matter
* Latitude
* Longitude

The data is being calculated by using the Air Quality Index.

**REGRESSION ANALYSIS**

In this measurable displaying, regression analysis is a factual procedure for assessing the connections among factors. It incorporates numerous systems for demonstrating and investigating a few factors, when the attention is on the relationship between a needy variable and at least one autonomous factors. All the more particularly, relapse investigation helps one see how the run of the mill estimation of the needy variable changes when any of the autonomous factors is shifted, while the other free factors are held settled. Most regularly, relapse investigation assesses the contingent desire of the needy variable given the autonomous factors, that is, the normal estimation of the reliant variable when the free factors are settled. Less normally, the emphasis is on a quantile, or other area parameter of the restrictive dispersion of the needy variable given the free factors. In all cases, the estimation target is an element of the free factors called the relapse work. In relapse examination, it is likewise important to portray the variety of the reliant variable around the relapse work which can be depicted by likelihood dissemination. A related however unmistakable approach is important condition investigation, which assesses the greatest (as opposed to normal) estimation of the needy variable for a given estimation of the free factor (roof line as opposed to focal line) with a specific end goal to recognize what estimation of the autonomous variable is essential yet not adequate for a given estimation of the reliant variable. The execution of relapse examination techniques by and by relies on upon the type of the information producing procedure, and how it identifies with the relapse approach being utilized. Since the genuine type of the information creating procedure is for the most part not known, relapse examination regularly depends to some degree on making presumptions about this procedure. These suppositions are now and again testable if an adequate amount of information is accessible.

Regression analysis is broadly utilized for expectation and determining, where its utilization has generous cover with the field of machine learning. Relapse investigation is additionally used to comprehend which among the autonomous factors are identified with the reliant variable, and to investigate the types of these connections. In limited conditions, relapse examination can be utilized to surmise causal connections between the free and ward factors.

In measurements, forecast is a piece of factual derivation. One specific way to deal with such deduction is known as prescient derivation, yet the expectation can be attempted inside any of the few ways to deal with factual induction. In fact, one depiction of insights is that it gives a method for exchanging information about a specimen of a populace to the entire populace, and to other related populaces, which is not really the same as forecast after some time. At the point when data is exchanged crosswise over time, regularly to particular focuses in time, the procedure is known as gauging.

Standard straight relapse models with standard estimation strategies make various suspicions about the indicator factors, the reaction factors and their relationship. Various expansions have been created that permit each of these suspicions to be casual (i.e. diminished to a weaker frame), and at times disposed of completely. A few strategies are sufficiently general that they can unwind various presumptions on the double, and in different cases this can be accomplished by consolidating diverse expansions. By and large these augmentations make the estimation method more intricate and tedious, and may likewise require more information so as to create a similarly exact model.

**TIME SERIES**

Time series involves strategies for breaking down time arrangement information with a specific end goal to separate important measurements and different attributes of the information. Time series is the utilization of a model to foresee future qualities in view of already watched values. While relapse investigation is frequently utilized so as to test speculations that the present estimations of at least one free time arrangement influence the present estimation of some other time arrangement, this kind of examination of time arrangement is not called "time arrangement examination", which concentrates on looking at estimations of a solitary time arrangement or numerous reliant time arrangement at various focuses in time.

Time series have a characteristic transient requesting. This sets aside a few minutes arrangement investigation unmistakable from cross-sectional reviews, in which there is no normal requesting of the Time arrangement examination is likewise particular from spatial information examination where the perceptions ordinarily identify with land areas. A stochastic model for a period arrangement will for the most part mirror the way that perceptions near one another in time will be more firmly related than perceptions facilitate separated. Furthermore, time arrangement models will regularly make utilization of the normal one-path requesting of time so that qualities for a given period will be communicated as getting somehow from past qualities, instead of from future qualities.

**PSEUDO CODE**

%python

import pandas as pd

import numpy as np

import matplotlib.pylab as plt

%matplotlib inline from matplotlib.pylab

import rcParams rcParams['figure.figsize'] = 15, 6

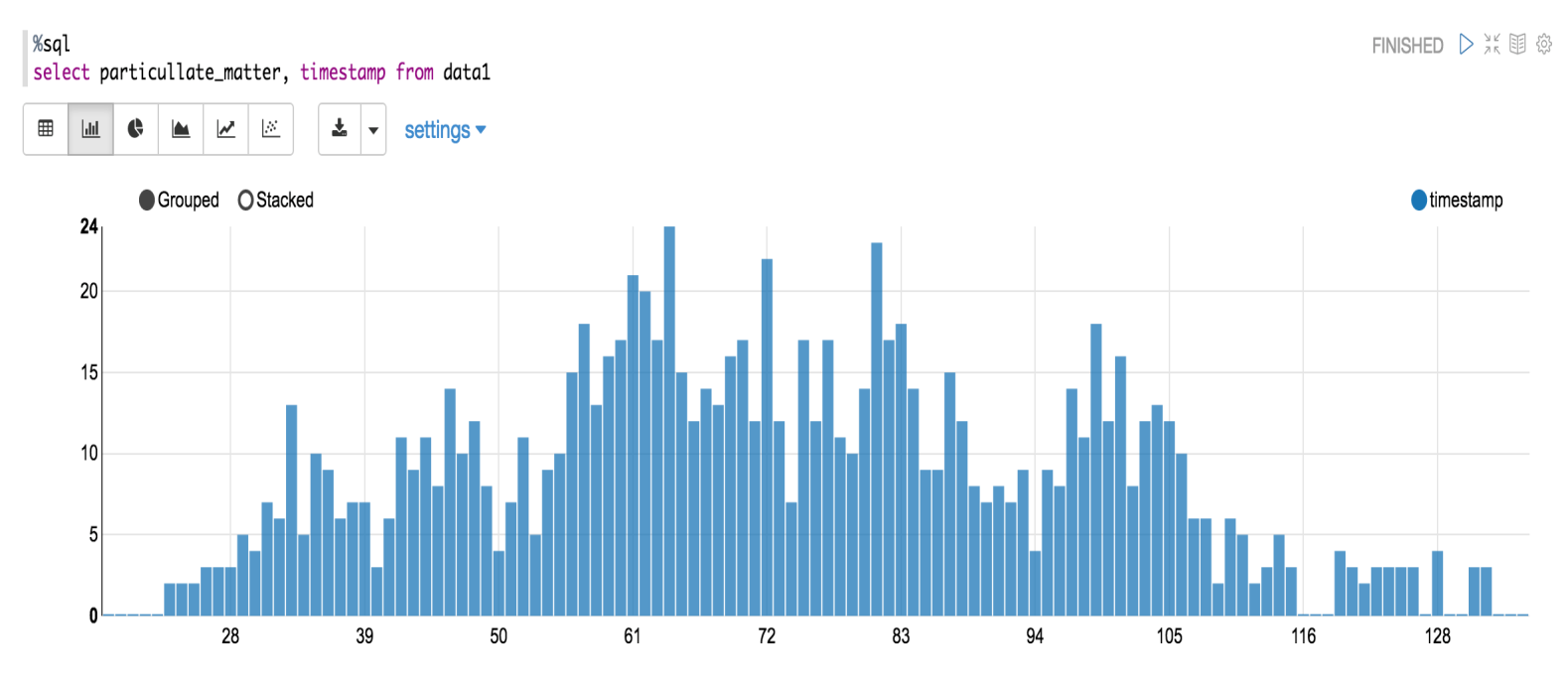
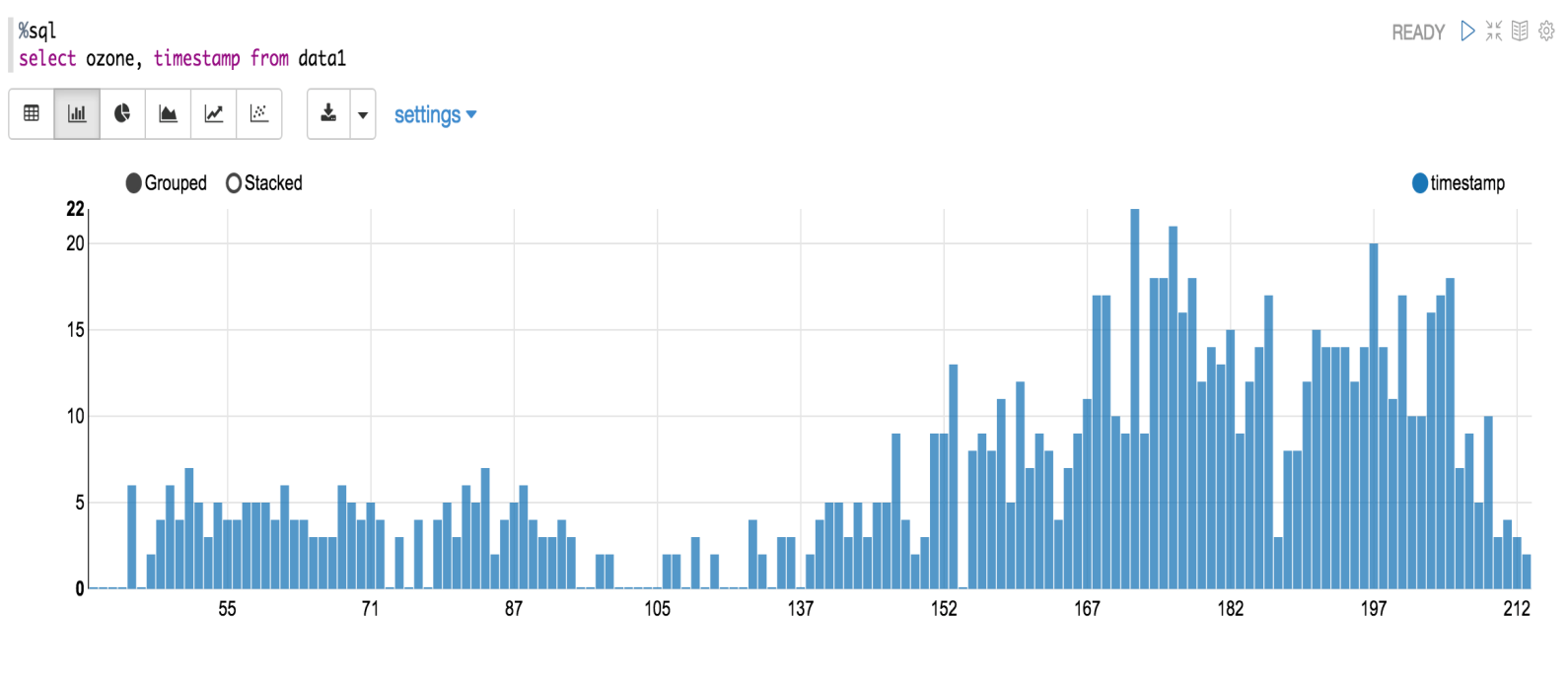
data = pd.read\_csv('Users/ZAhnad/Downloads/pollution/\*')

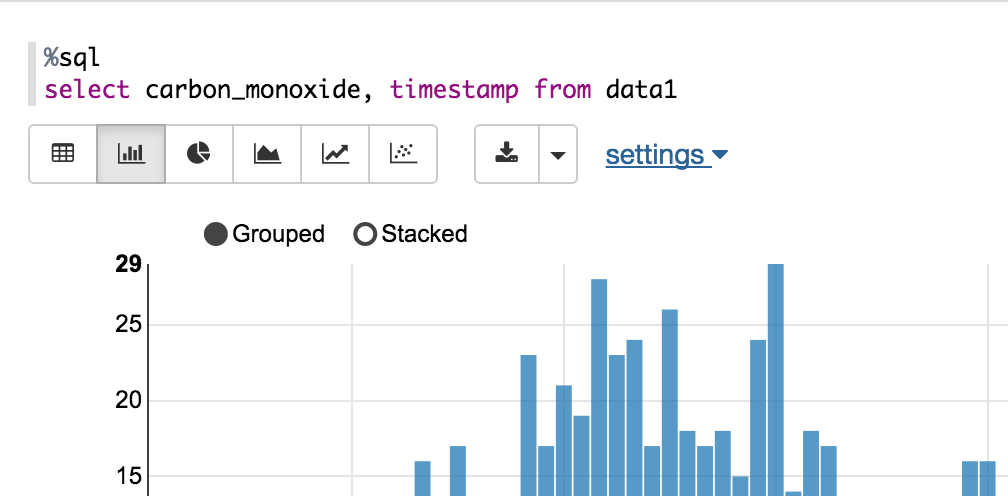
print data.head()

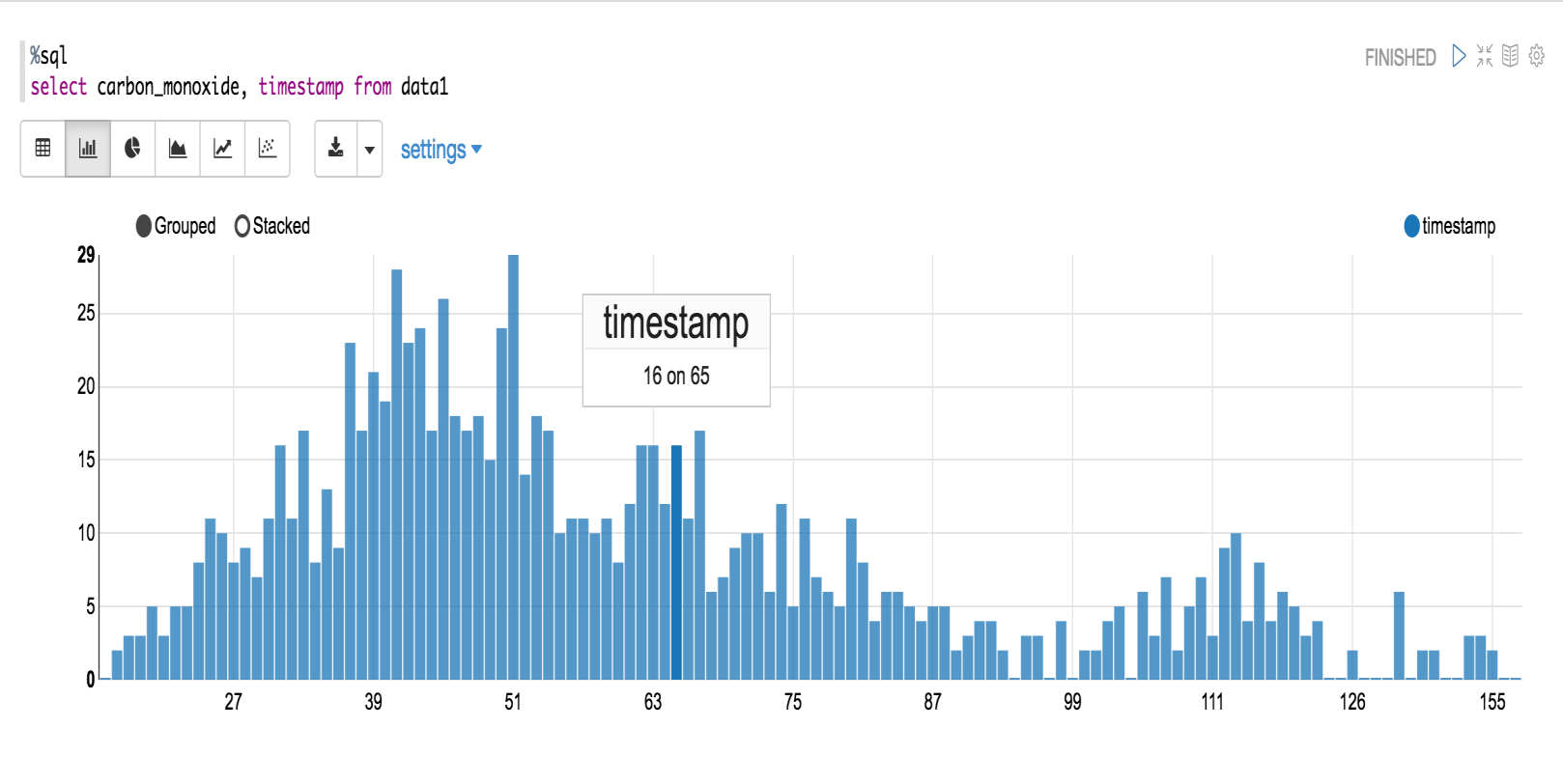
%sql

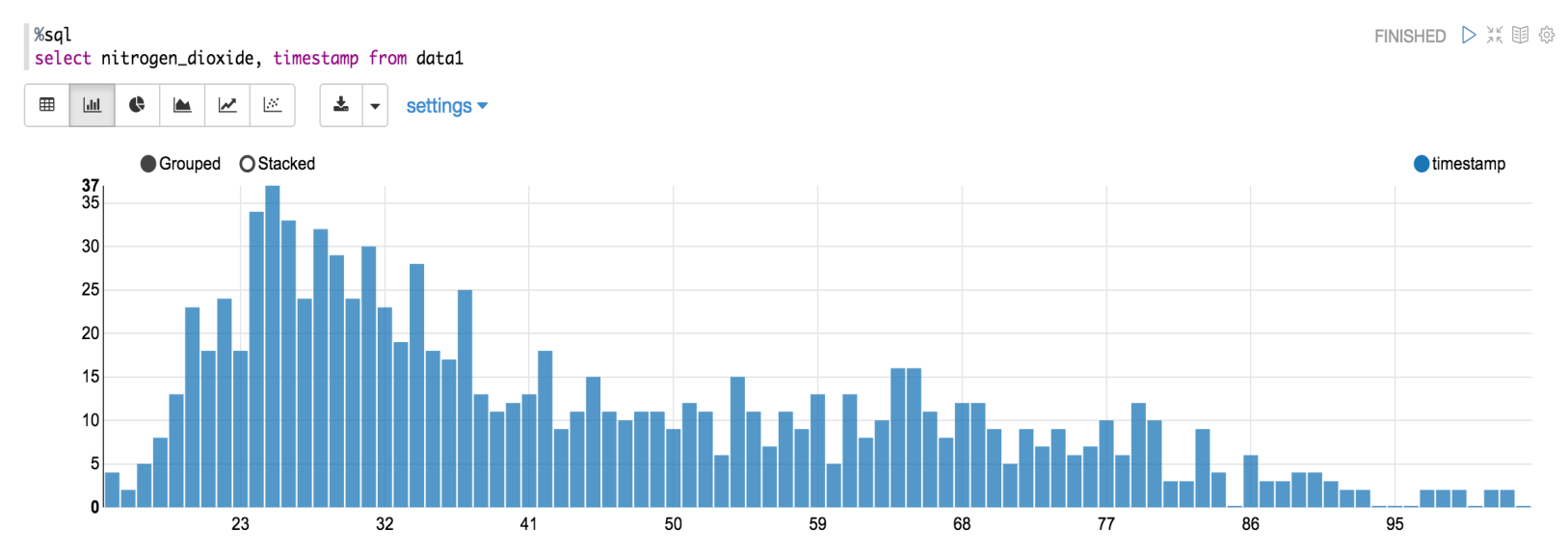
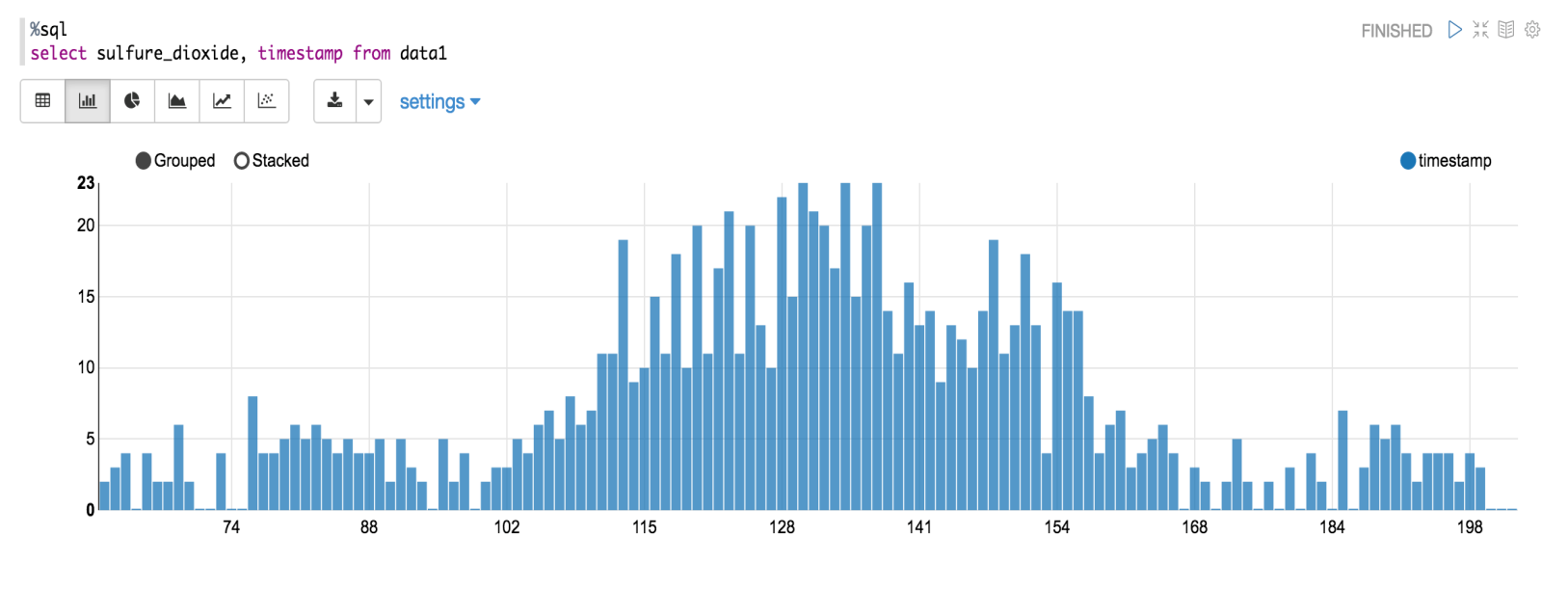
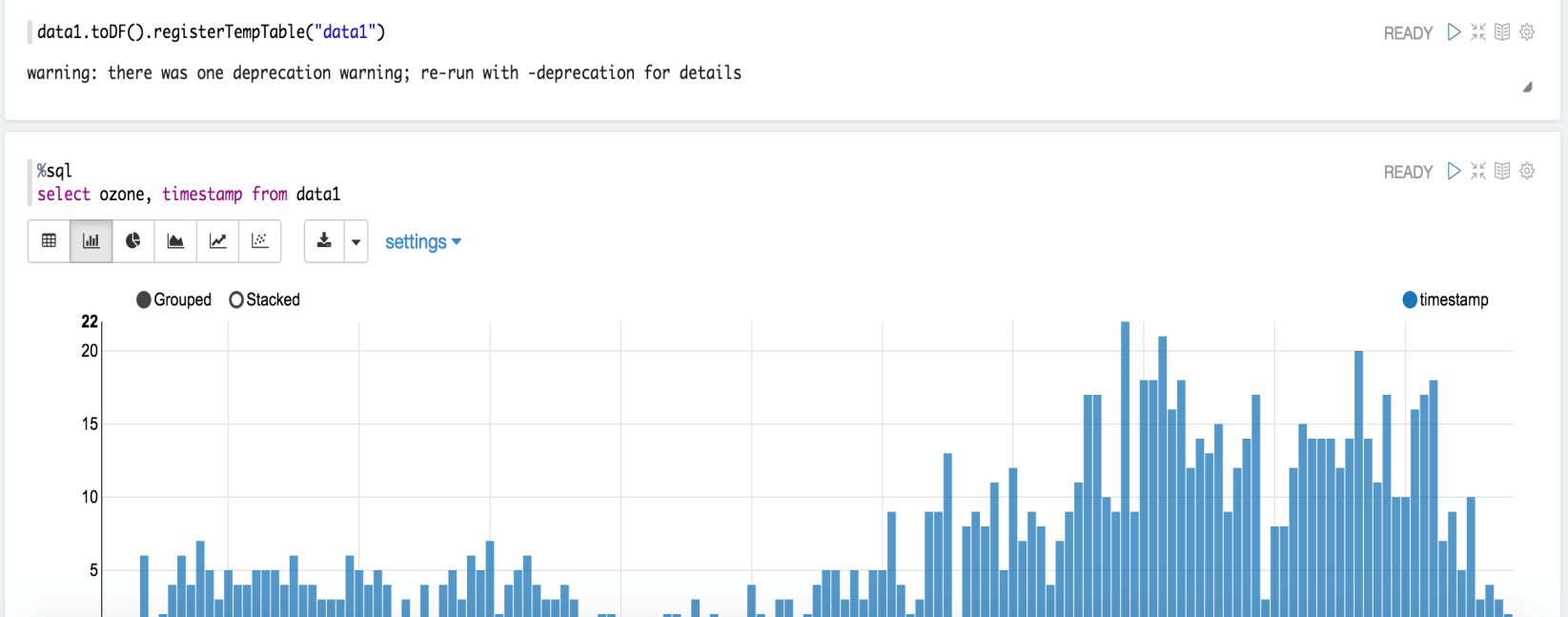
select carbon\_monoxide, timestamp from data1

**ZEPELLIN GRAPHS**









**REFERENCES**

Berkowicz, R., Palmgren, F., Hertel, O., & Vignati, E. (1996). Using measurements of air pollution in streets for evaluation of urban air quality—meterological analysis and model calculations. Science of the total environment, 189, 259-265.

FitzHugh, T. W., & Mackay, D. S. (2000). Impacts of input parameter spatial aggregation on an agricultural nonpoint source pollution model. Journal of hydrology, 236(1), 35-53.

In this measurable displaying, regression analysis is a factual procedure for assessing the connections among factors. It incorporates numerous systems for demonstrating and investigating a few factors, when the attention is on the relationship between a needy variable and at least one autonomous factor. All the more particularly, relapse investigation helps one see how the run of the mill estimation of the needy variable changes when any of the autonomous factors is shifted, while the other free factors are held settled. Most regularly, relapse investigation assesses the contingent desire of the needy variable given the autonomous factors, that is, the normal estimation of the variable when the free factors are settled. Less normally, the emphasis is on a quantile, or other area parameter of the restrictive dispersion of the needy variable given the free factors. In all cases, the estimation target is an element of the free factors called the relapse work. In relapse examination, it is likewise important to portray the variety of the variable around the relapse work which can be depicted by likelihood dissemination. A related however unmistakable approach is important condition investigation, which assesses the greatest (as opposed to normal) estimation of the needy variable for a given estimation of the free factor (roof line as opposed to focal line) with a specific end goal to recognize what estimation of the autonomous variable is essential yet not adequate for a given estimation of the reliant variable. The execution of relapse examination techniques by and by relies on upon the type of the information producing procedure, and how it identifies with the relapse approach being utilized. Since the genuine type of the information creating procedure is for the most part not known, relapse examination regularly depends to some degree on making presumptions about this procedure. These suppositions are now and again testable if an adequate amount of information is accessible. Regression analysis is broadly utilized for expectation and determining, where its utilization has generous cover with the field of machine learning. Relapse investigation is additionally used to comprehend which among the autonomous factors are identified with the variable, and to investigate the types of these connections. In limited conditions, relapse examination can be utilized to surmise causal connections between the free and ward factors. In measurements, forecast is a piece of factual derivation. One specific way to deal with such deduction is known as prescient derivation, yet the expectation can be attempted inside any of the few ways to deal with factual induction. In fact, one depiction of insights is that it gives a method for exchanging information about a specimen of a populace to the entire populace, and to other related populaces, which is not really the same as forecast after some time. At the point when data is exchanged crosswise over time, regularly to particular focuses in time, the procedure is known as gauging. In this measurable displaying, regression analysis is a factual procedure for assessing the connections among factors. It incorporates numerous systems for demonstrating and investigating a few factors, when the attention is on the relationship between a needy variable and at least one autonomous factor. All the more particularly, relapse investigation helps one see how the run of the mill estimation of the needy variable changes when any of the autonomous factors is shifted, while the other free factors are held settled. Most regularly, relapse investigation assesses the contingent desire of the needy variable given the autonomous factors, that is, the normal estimation of the variable when the free factors are settled. Less normally, the emphasis is on a quantile, or other area parameter of the restrictive dispersion of the needy variable given the free factors. In all cases, the estimation target is an element of the free factors called the relapse work. In relapse examination, it is likewise important to portray the variety of the variable around the relapse work which can be depicted by likelihood dissemination. A related however unmistakable approach is important condition investigation, which assesses the greatest (as opposed to normal) estimation of the needy variable for a given estimation of the free factor (roof line as opposed to focal line) with a specific end goal to recognize what estimation of the autonomous variable is essential yet not adequate for a given estimation of the reliant variable. The execution of relapse examination techniques by and by relies on upon the type of the information producing procedure, and how it identifies with the relapse approach being utilized. Since the genuine type of the information creating procedure is for the most part not known, relapse examination regularly depends to some degree on making presumptions about this procedure. These suppositions are now and again testable if an adequate amount of information is accessible. Regression analysis is broadly utilized for expectation and determining, where its utilization has generous cover with the field of machine learning. Relapse investigation is additionally used to comprehend which among the autonomous factors are identified with the variable, and to investigate the types of these connections. In limited conditions, relapse examination can be utilized to surmise causal connections between the free and ward factors. In measurements, forecast is a piece of factual derivation. One specific way to deal with such deduction is known as prescient derivation, yet the expectation can be attempted inside any of the few ways to deal with factual induction. In fact, one depiction of insights is that it gives a method for exchanging information about a specimen of a populace to the entire populace, and to other related populaces, which is not really the same as forecast after some time. At the point when data is exchanged crosswise over time, regularly to particular focuses in time, the procedure is known as gauging.