

New Zealand

Earthquake study

Nestor Pereira | STAT 279 – Data science | 10/16/2019

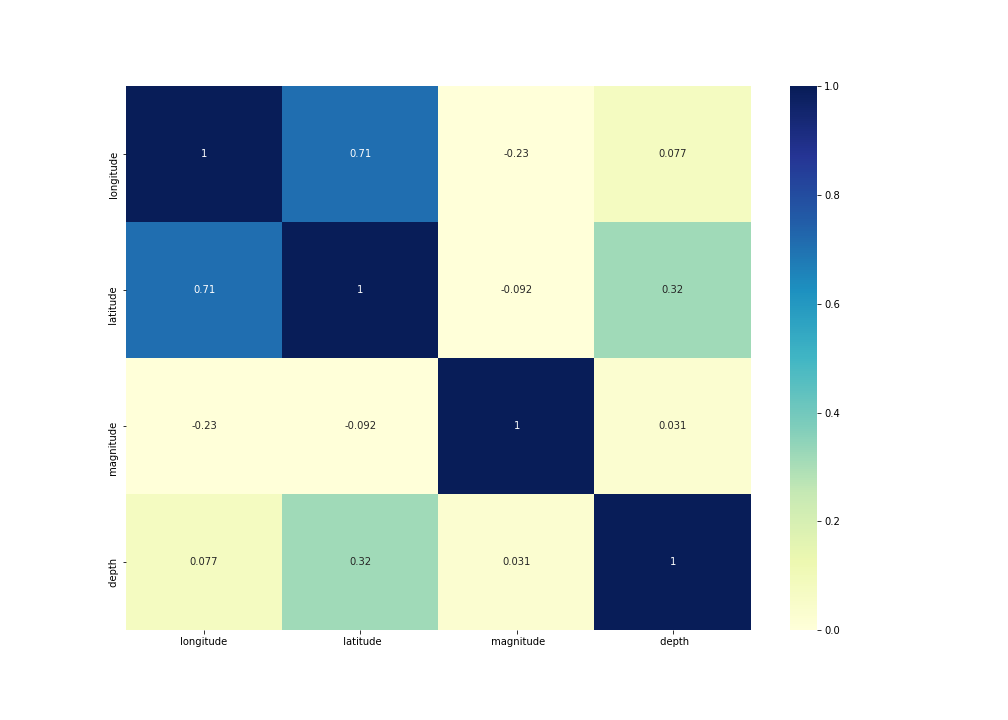
For our study we choose the region of wellington for its frequent number of earthquakes. I was born in June 14 so we choose to study the earthquake between June 14 2019 and July 13 of 2019.

# How many detectable earthquakes were there in your data?

After cleaning the data and making sure the only eventtype chosen were earthquake we obtain a data frame with 367 observations.

# Make a correlation matrix using proc corr of the variables. What is the relationship between magnitude and depth in your data?

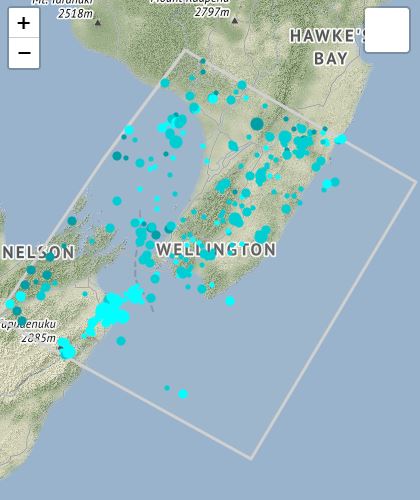
We can calculate a covariance matrix which shows the correlation of all the variables and plot it as a heatmap.



We can see that the relation between magnitude and depth is 0.031, meaning almost zero. Then we can say there is no correlation between these two features.

# Is there a correlation between latitude and longitude for this data?

Looking at the heatmap of correlation from before we can see that there is a strong correlation of 0.71. We can understand this correlation looking at the map of New Zealand.

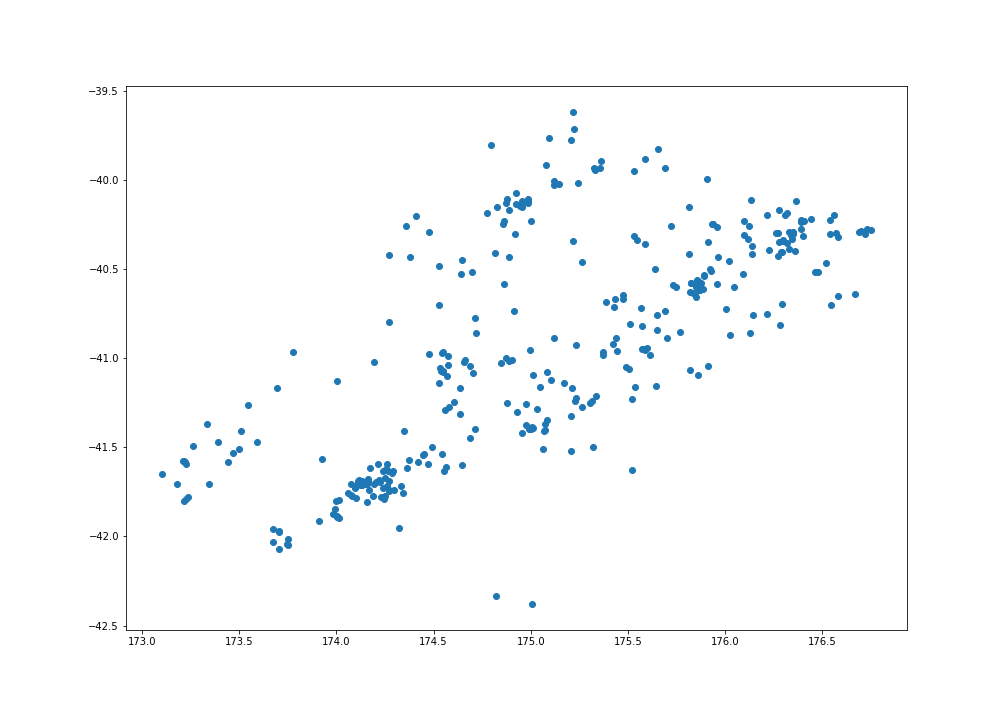


We can see that most of the earthquake follow the direction of the land. Earthquake in the south part of our study area happen on the west side and the earthquake that happen on the north side of the area happen on the east side.

This means the more south (smaller latitude) the earthquake happens the more west ( smaller longitude) it will happen and viceversa.

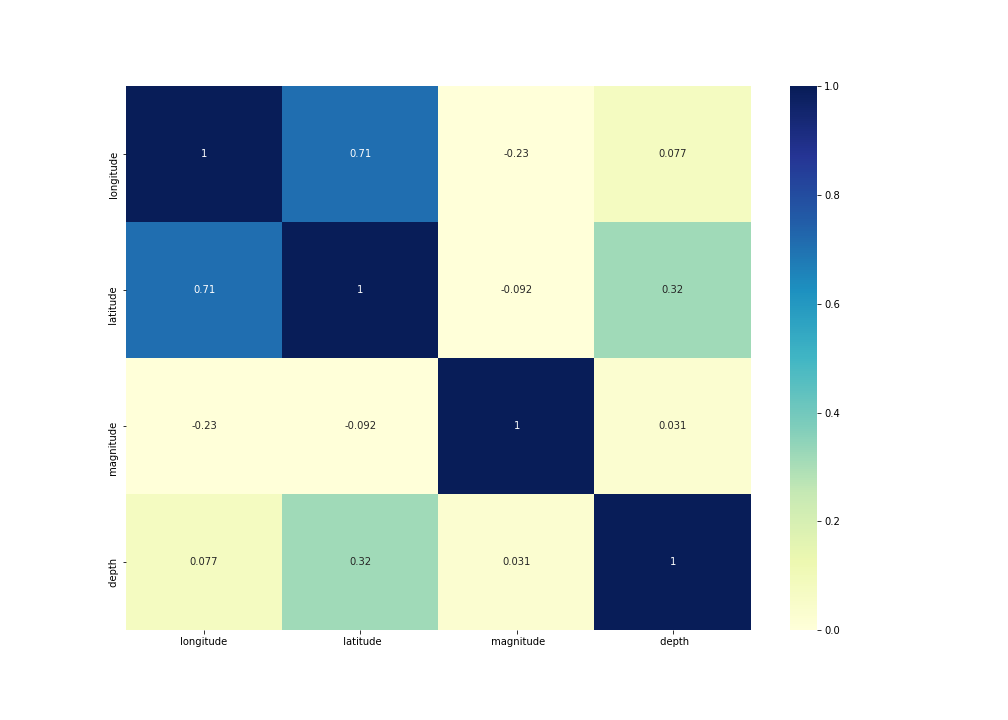
# 7. Make a plot of the latitude (y-axis) against longitude (x-axis) of the earthquakes and describe the plot

I’m going to skip to question 7 since its very related to question 3. As described before, the smaller the latitude the smaller the longitude as show in the next plot. Also, that explains the positive strong 0.71 correlation between the variables.



# Is there a correlation between latitude and magnitude for this data? If so, what might this mean for the two main islands of New Zealand (North Island and South Island)?

Let’s look at the heatmap of correlations again.



We can find that there is a negative very small latitude between magnitude and how south the earthquake happens. Meaning there is not really a correlation between these variables. Answering to the second question, based on theses results, earthquakes in the south island are just as weak or strong as the ones in the northern island.

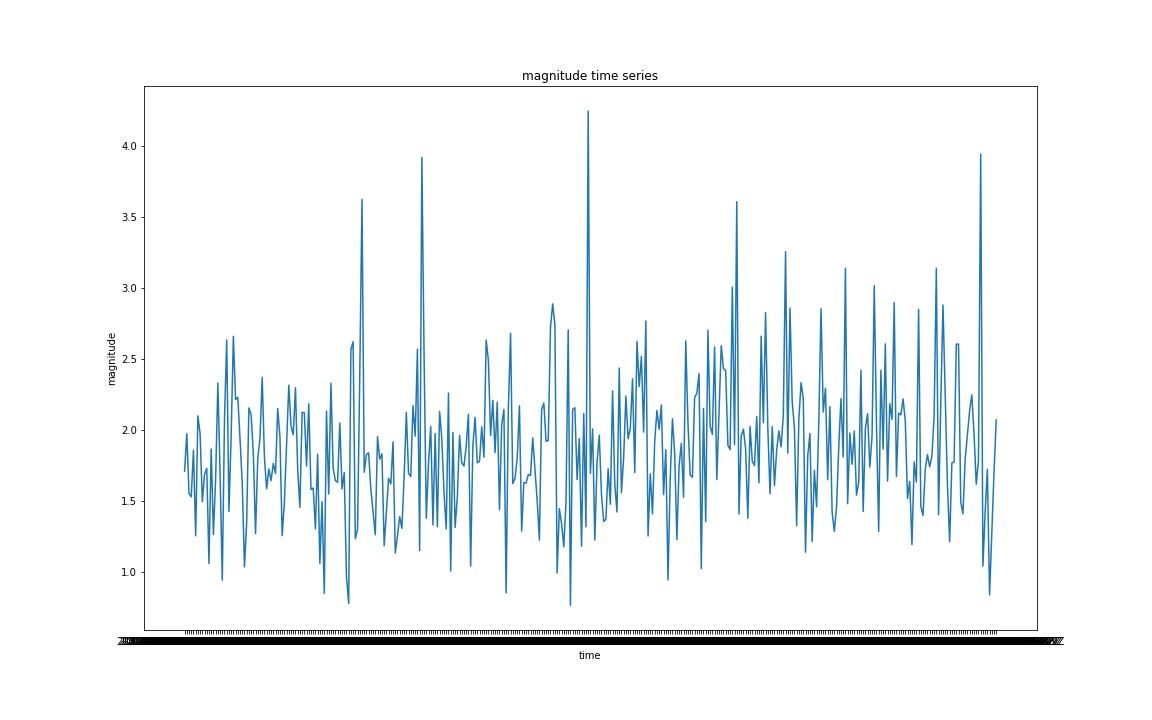
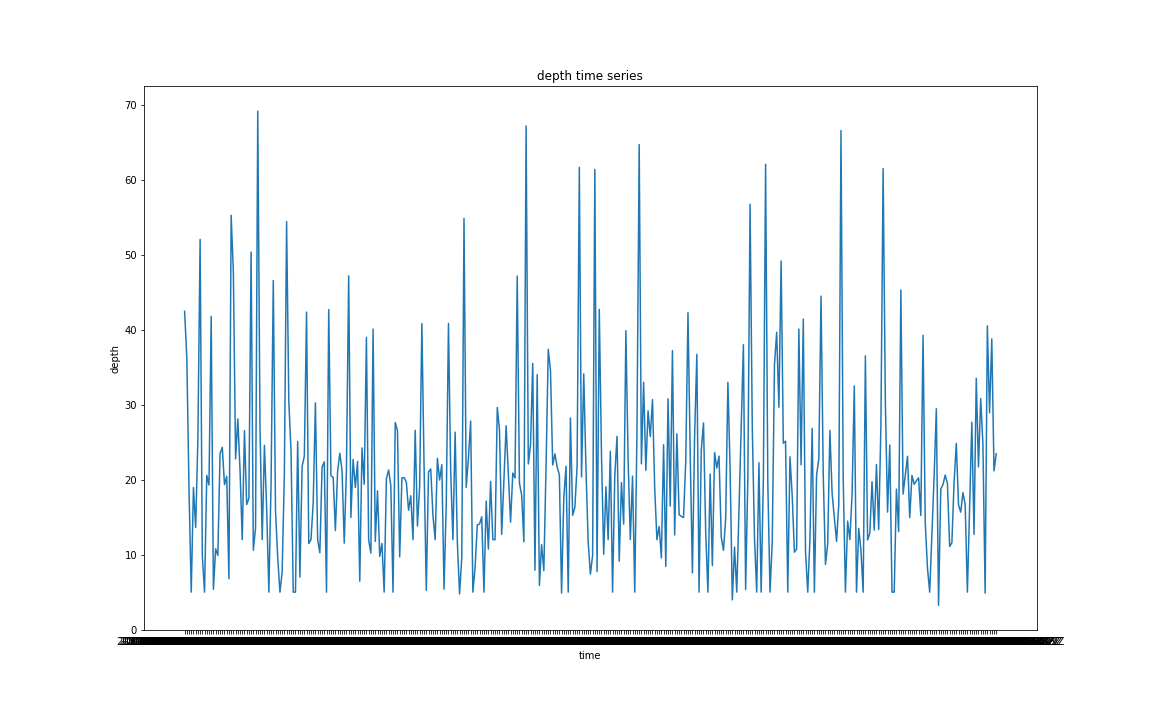
# Describe missingness in the data. For the entire dataset, how many observations had at least one missing value for the variables of interest? What percentage of the data is missing.

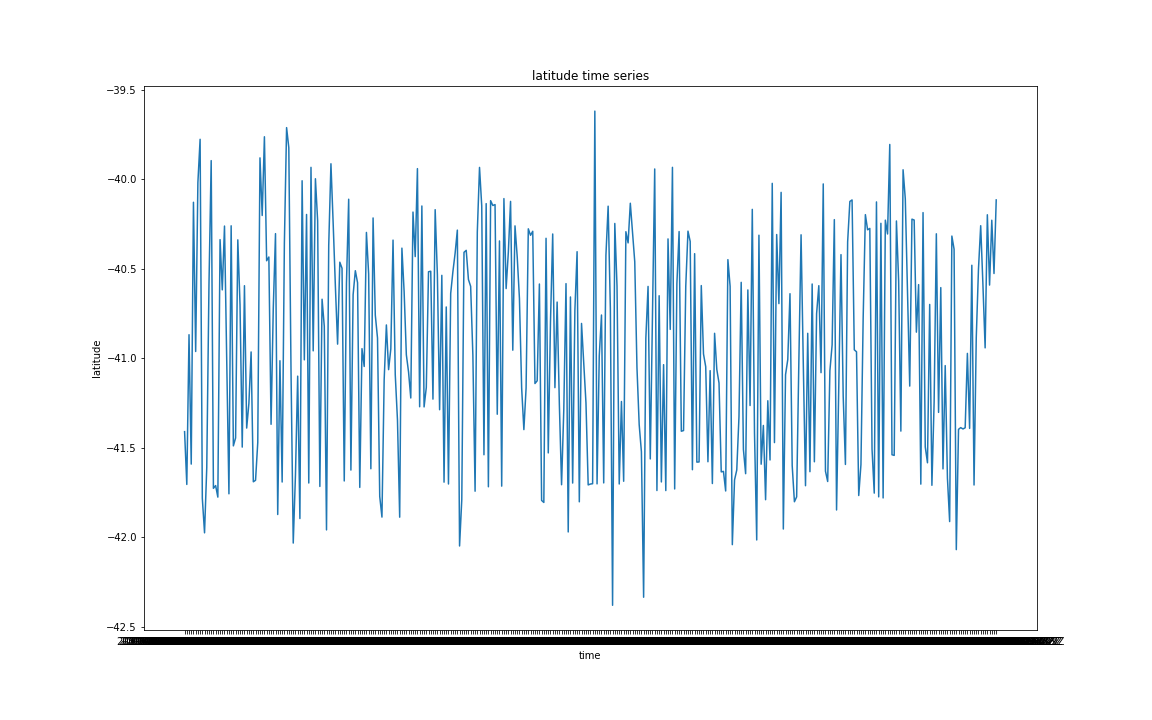
There are 324 earthquake who have missing values in depthtype. There are also 3 observation who has 3 missing values for the variable evaluation status, meaning that the earthquake has not been confirmed.

In total there are 324 observation with at least one missing values out of the 370 observation.

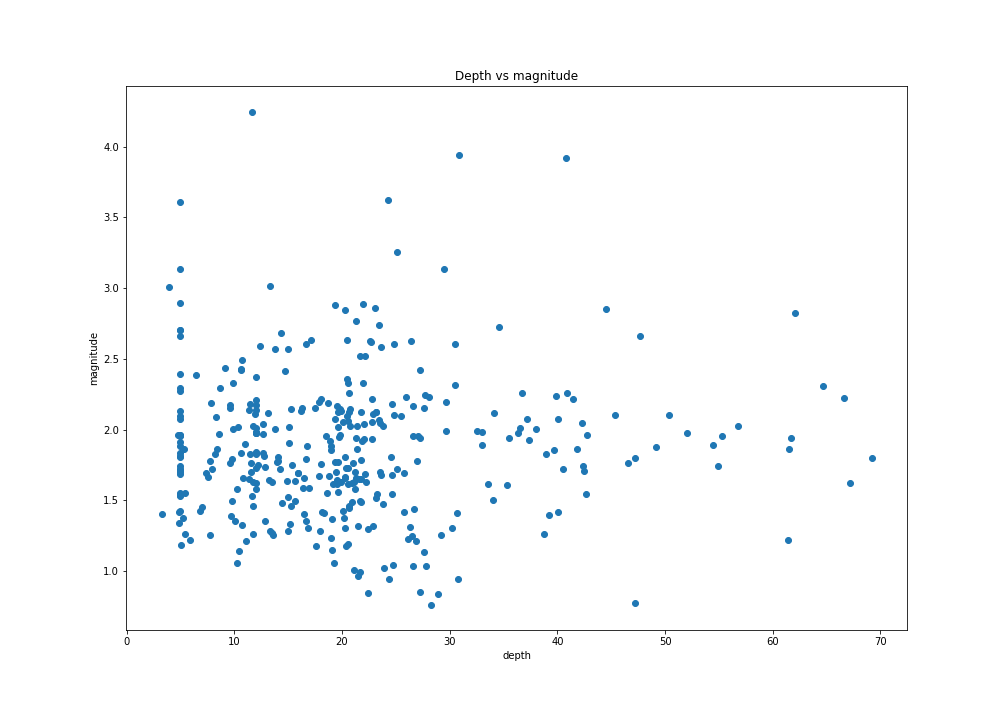
If we only consider the variables origintime, longitude, latitude, magnitude and depth, there are no missing values. If we considered the whole dataset, there is a percentage of 5% of the data is missing for the 324 observations out of the 370.

Make three plots which are time series of the earthquakes depths, magnitudes, and latitudes. (Here, you can plot the variable of interest against time.)





# 8.Make a plot of the magnitude (y-axis) against depth (x-axs) of the earthquakes and describe the plot.



Looking at the plot, I don’t see a correlation between these two variables, but just to make sure we go back to question 2 where we found that the correlation between depth and magnitude was 0.031, almost zero. That proves our point and correspond to the data plotted here.

# 9.Make a variable which is the time between successive earthquakes. Make a histogram of the times between earthquakes and describe the distribution

I wasn’t able to convert the variable origintime into datetime format so I could play with it.

The idea is to create two different vector where one is an earthquake less. Substract them and obtain the time difference between all the earthquakes in a vector. The just histogram plot that vector and try to the how often vector occur. Maybe that distribution was going to be a normal distribution.

I’ll leave the code here if that helps:

temp=df.loc[1:,['origintime']]

tempo=df.loc[:-1,['origintime']]

time=tempo-temp

plt.figure(figsize=(14,10))

np.histogram(time,bins=30)

plt.show()

ANNEX

import seaborn as sns

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import datetime

df=pd.read\_csv("earthquakes.csv")

df=df[df["eventtype"]=="earthquake"]

df=df.loc[:,['origintime','longitude',' latitude',' magnitude',' depth']]

df.to\_csv("earthquake\_clean.csv")

# 1. How many detectable earthquakes were there in your data?

df.shape[0]

# 2. Make a correlation matrix using proc corr of the variables. What is the relationship between magnitude and depth in your data?

cor=df.corr(method="pearson")

plt.figure(figsize=(14,10))

sns.heatmap(cor,vmin=0, vmax=1,annot=True, cmap="YlGnBu")

plt.savefig("correlationmatrix")

plt.show()

print("What is the relationship between magnitude and depth in your data?")

print(cor.iloc[2,3])

#3. Is there a correlation between latitude and longitude for this data? How can you interpret a correlation between these variables?

print("What is the relationship between latitude and longitude in your data?")

print(cor.iloc[1,0])

# 4. Is there a correlation between latitude and magnitude for this data? If so, what might this mean for the two main islands of New Zealand (North Island and South Island)?

print("Is there a correlation between latitude and magnitude for this data?")

print(cor.iloc[1,2])

print("Not a strong one, there is a very small negative correlation. Notice that the smaller the latitude the more south we are. This small correlation would mean that the smaller the latitude or the more south we are the higher the magnitude, but it's too small to be significant and considered as a strong relevant correlation")

plt.figure(figsize=(14,10))

plt.scatter(df.as\_matrix([' latitude']),df.as\_matrix([' magnitude']))

plt.show()

# 5. Describe missingness in the data. For the entire dataset, how many observations had at least one missing value for the variables of interest? What percentage of the data is missing

df=pd.read\_csv("earthquakes.csv")

df.isnull().sum()

df.loc[:,["depthtype"]]

a=df[df["depthtype"]!="operator assigned"]

a.info()

# there are 324 earthquake who have NA in depthtype. There are also 3 observation who has 3 missing values for the variable evaluation status, meaning that the earthquake hasent been confirmed.

# in total there are 324 observation with at least one missing values out of the 370 observation.

# 6. Make three plots which are time series of the earthquakes depths, magnitudes, and latitudes. (Here,

#you can plot the variable of interest against time.) Make the points connected by lines, as is typically

#done in time series plots

df=pd.read\_csv("earthquakes.csv")

df=df[df["eventtype"]=="earthquake"]

df=df.loc[:,['origintime','longitude',' latitude',' magnitude',' depth']]

df.to\_csv("earthquake\_clean.csv")

b=df.as\_matrix(['origintime'])

plt.figure(figsize=(16,10))

plt.title("depth time series")

plt.xlabel("time")

plt.ylabel("depth")

plt.plot(df.as\_matrix(['origintime']),df.as\_matrix([' depth']))

plt.savefig('depthtimeseries')

plt.show()

plt.figure(figsize=(16,10))

plt.title("magnitude time series")

plt.xlabel("time")

plt.ylabel("magnitude")

plt.plot(df.as\_matrix(['origintime']),df.as\_matrix([' magnitude']))

plt.savefig('magnitudetimeseries')

plt.show()

plt.figure(figsize=(16,10))

plt.title("latitude time series")

plt.xlabel("time")

plt.ylabel("latitude")

plt.plot(df.as\_matrix(['origintime']),df.as\_matrix([' latitude']))

plt.savefig('latitudetimeseries')

plt.show()

# 7. Make a plot of the latitude (y-axis) against longitude (x-axis) of the earthquakes and describe the plot

plt.figure(figsize=(14,10))

plt.scatter(df.as\_matrix(['longitude']),df.as\_matrix([' latitude']))

plt.savefig('latencyvslongitude')

plt.show()

# 8. Make a plot of the magnitude (y-axis) against depth (x-axis) of the earthquakes and describe the plot.

plt.figure(figsize=(14,10))

plt.scatter(df.as\_matrix([' depth']),df.as\_matrix([' magnitude']))

plt.xlabel("depth")

plt.ylabel("magnitude")

plt.title("Depth vs magnitude")

plt.savefig('depthvsmagnitude')

plt.show()

#9. Make a variable which is the time between successive earthquakes. Make a histogram of the times

#between earthquakes and describe the distribution.

temp=df.loc[1:,['origintime']]

tempo=df.loc[:368,['origintime']]

time=tempo-temp

plt.figure(figsize=(14,10))

np.histogram(time,bins=30)

plt.show()