**KIIT University , Bhubaneswar**

**Data Science Using Python**

**Tedx Talks DATA ANALYSIS**

**Training Project Report**

**A report of four weeks Industrial training**

**At**

**Webtek Labs**

**CANDIDATE’S DECLARATION**

**I hereby declare that I have undertaken an industrial training at “WEBTEK LABS” during a period from 16th May to 15th June in partial fulfilment of requirements for the award of degree of B.E (Computer science) .**

**Signature of the student**

**The four weeks industrial training Viva-voice examination of**

**Has been held on and accepted.**

**CERTIFICATE OF APPROVAL**

The project “**TeDx Talks Data Anlaysis**” made by **Vaibhav Kumar** is hereby approved as a creditable study for the Bachelor of Engineering in Computer Science and presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned this project only for the purpose for which it is submitted.

**MS. MOUSITA DHAR**

**(Project In-charge)**

**ACKNOWLEDGEMENT**

I would like to express our special thanks of gratitude to our trainer MS.MOUSITA DHAR who gave me the golden opportunity to do this wonderful project on the topic **TEDX TALKS DATA ANALYSIS BY DATA SCIENCE USING PYTHON** which also helped us in doing a lot of Research and I came to know about so many new things. I am really thankful to her.  
Finally, I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

**Signature of project mentor**

**1. INTRODUCTION**

**1.1 PYTHON**

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

* **Python is Interpreted** − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* **Python is Interactive** − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
* **Python is Object-Oriented** − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
* **Python is a Beginner's Language** − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

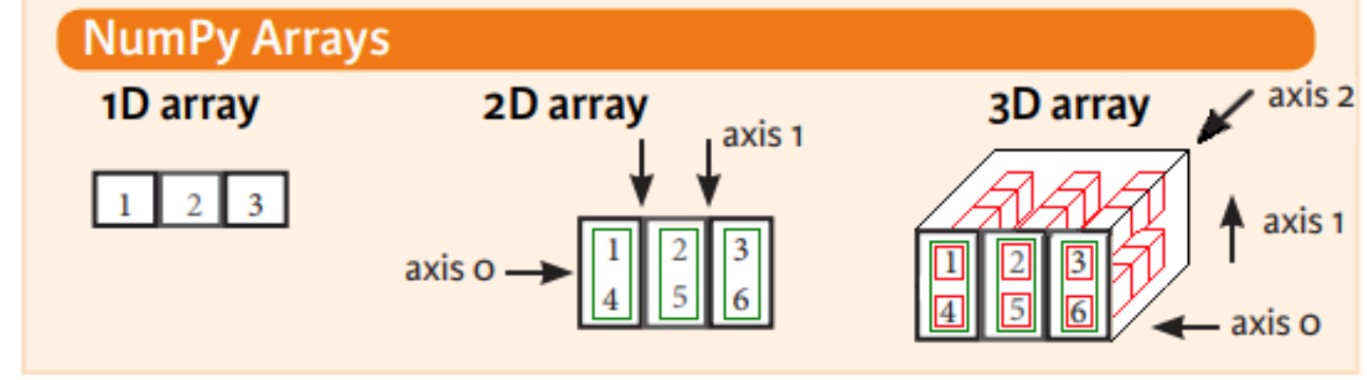
PYTHON FEATURES

* **Easy** to Learn and Use. Python is **easy** to learn and use
* Expressive Language
* Interpreted Language
* Cross-**platform** Language
* Free and Open Source
* **Object-Oriented** Language
* Extensible
* Large Standard Library
* **APPLICATIONS OF PYTHON**
* Web and internet development
* Scientific and numeric computing
* Data Analysis
* Desktop GUIs
* Machine Learning
* Data visualization
* Game Development
* Software Development
* Business Application
  1. **ANACONDA**

**Anaconda** is a free and open distribution of Python programming languages for data science and machine learning related applications (large-scale data processing, predictive analytics, scientific computing), that aims to simplify package management and deployment. Package versions are managed by the package management system *conda*. Conda is an open source, cross platform, language-agnostic package manager and environment management system that installs, runs, and updates packages and their dependencies. The Anaconda distribution is used by over 6 million users, and it includes more than 250 popular data science packages suitable for Windows, Linux, and MacOS.

* 1. **PYTHON PACKAGES**
* **NUMPY**
* NumPy is the fundamental package for scientific computing with Python. It contains among other things:
* a powerful N-dimensional array object
* sophisticated (broadcasting) functions
* tools for integrating C/C++ and Fortran code
* useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

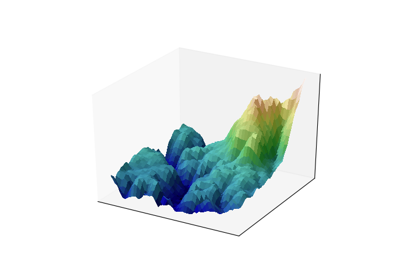
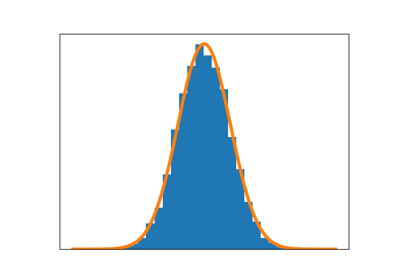


* **Matplotlib**

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when

Combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc., via an object oriented interface or via a set of functions familiar to MATLAB users.

* **Pandas**

*Pandas* is an open source, BSD-licensed library providing high-performance, easy- to-use data structures and data analysis tools for the *Python* programming language. Pandas library is well suited for data manipulation and analysis using python. In particular, it offers data structures and operations for manipulating numerical tables and time series.

* **Scikit-learn**

Scikit-learn provide machine learning libraries for python some of the features of Scikit- learn includes:

* Simple and efficient tools for data mining and data analysis
* Accessible to everybody, and reusable in various contexts
* Built on NumPy, SciPy, and matplotlib
* Open source, commercially usable - BSD license

**TRAINING WORK UNDERTAKEN**

* **COLLECTING DATA FROM KAGGLE**

**Kaggle** is a platform for predictive modelling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. This crowd sourcing approach relies on the fact that there are countless strategies that can be applied to any predictive modelling task and it is impossible to know beforehand which technique or analyst will be

most effective.

On 8 March 2017, Google announced that they were acquiring Kaggle. They will join the Google Cloud team and continue to be a distinct brand. In January 2018, Booz Allen and Kaggle launched Data Science Bowl, a machine learning competition to analyse cell images and identify nuclei.

* **DATA SCIENCE**

**Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, structured and unstructured, similar to data mining. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyse actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

When Harvard Business Review called it "The Sexiest Job of the 21st Century"the term became a buzzword, and is now often applied to business analytics, business intelligence, predictive modelling, or any arbitrary use of data, or used as a glamorized term for statistics. In many cases, earlier approaches and solutions are now simply rebranded as

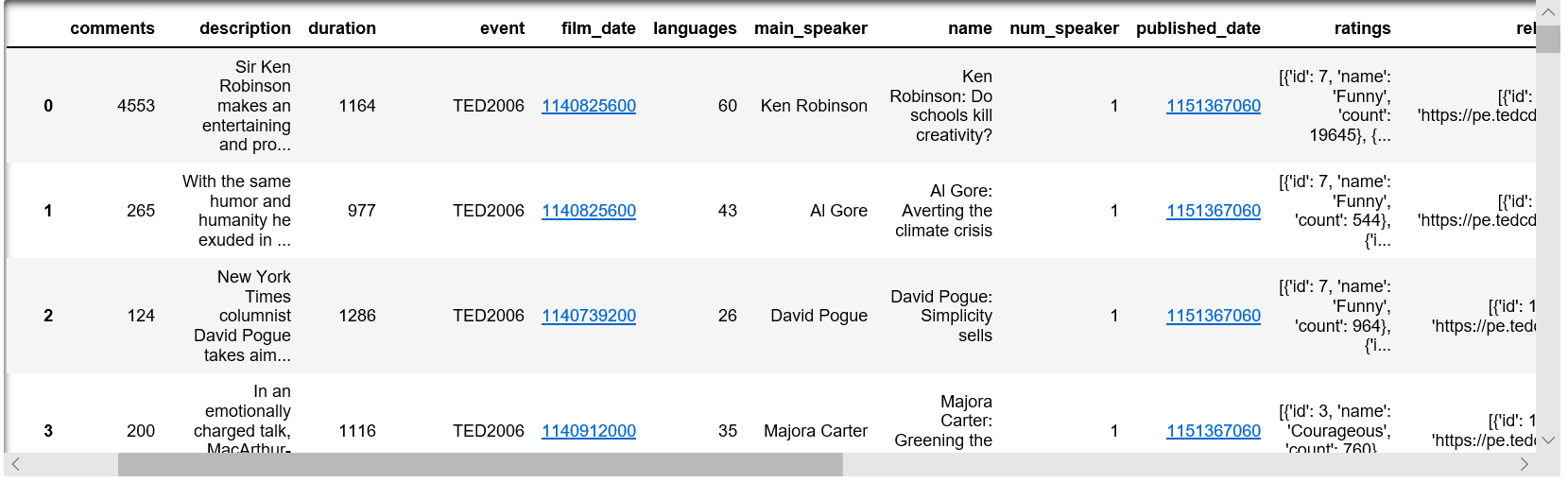
"Data science" to be more attractive, which can cause the term

"Data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness." While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents. Because of the current popularity

of this term, there are many "advocacy efforts" surrounding the field. To its discredit, however, many data science and big data projects fail to deliver useful results, often as a result of poor management and utilization of resources.

* **DATASET**

The dataset contains 2550 rows and 17 columns.



**Source code and output:**

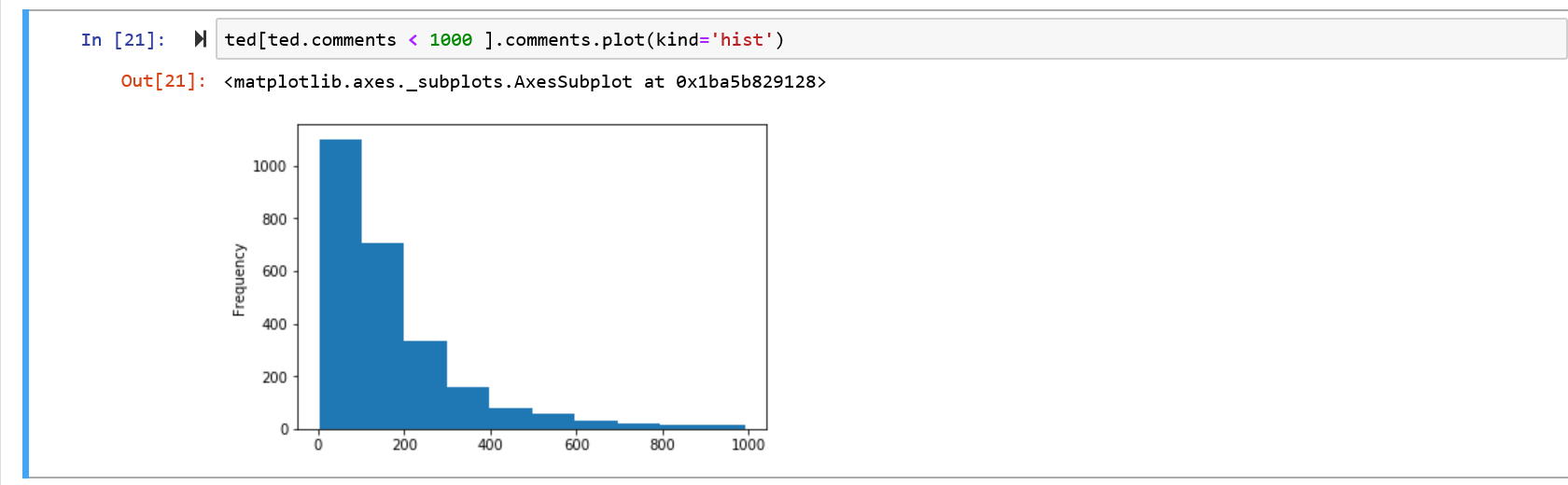
**Importing modules and loading of the dataset:**

First, I install all the libraries that I will use in our application. I install all the libraries in the first part because the algorithms I will use later and the analysis I will make more clearly will be done.Furthurmore, I have investigated the data, presented some visualization and analysed features.  After that we are uploading our data set to the data variable using the **read\_csv** function in the pandas library. The **head ()** function will show 5 values by-default.

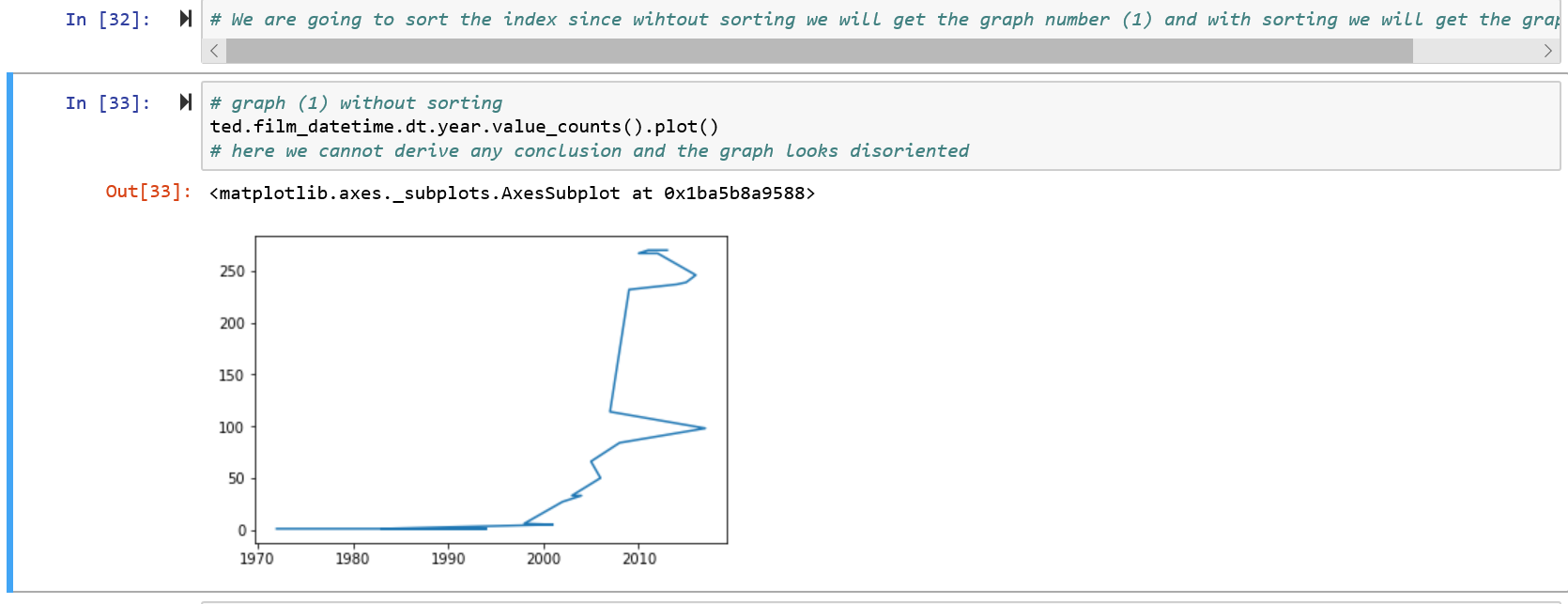


**Data Analysis:**

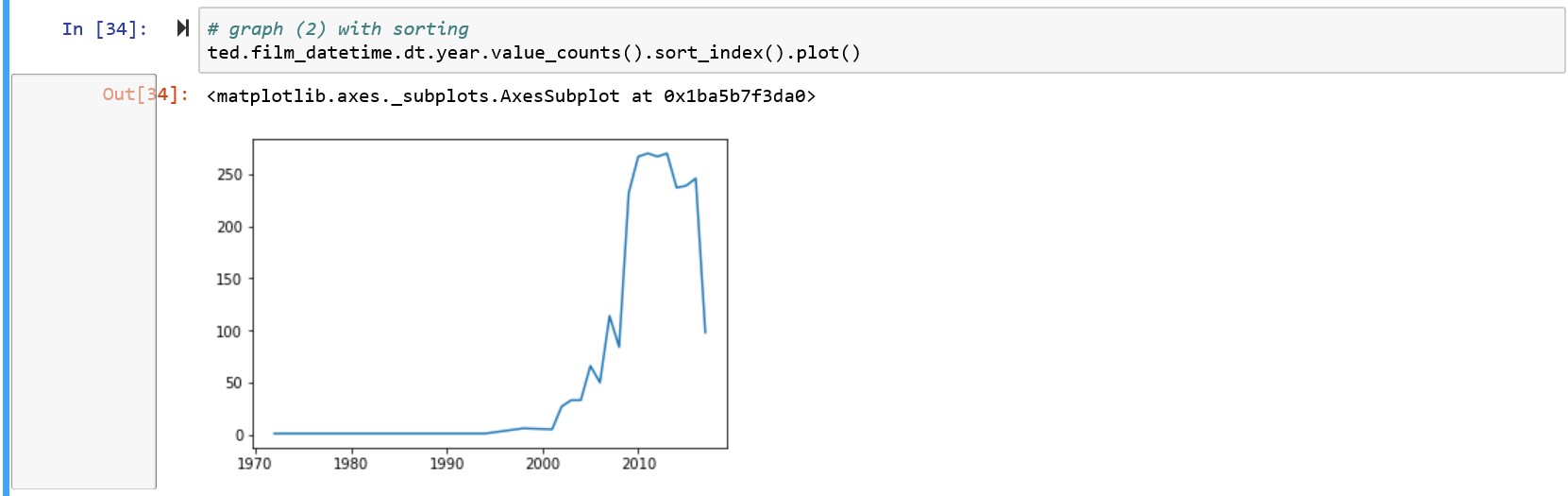
* 1. **Histogram plot of comments less than 1000 .**

****

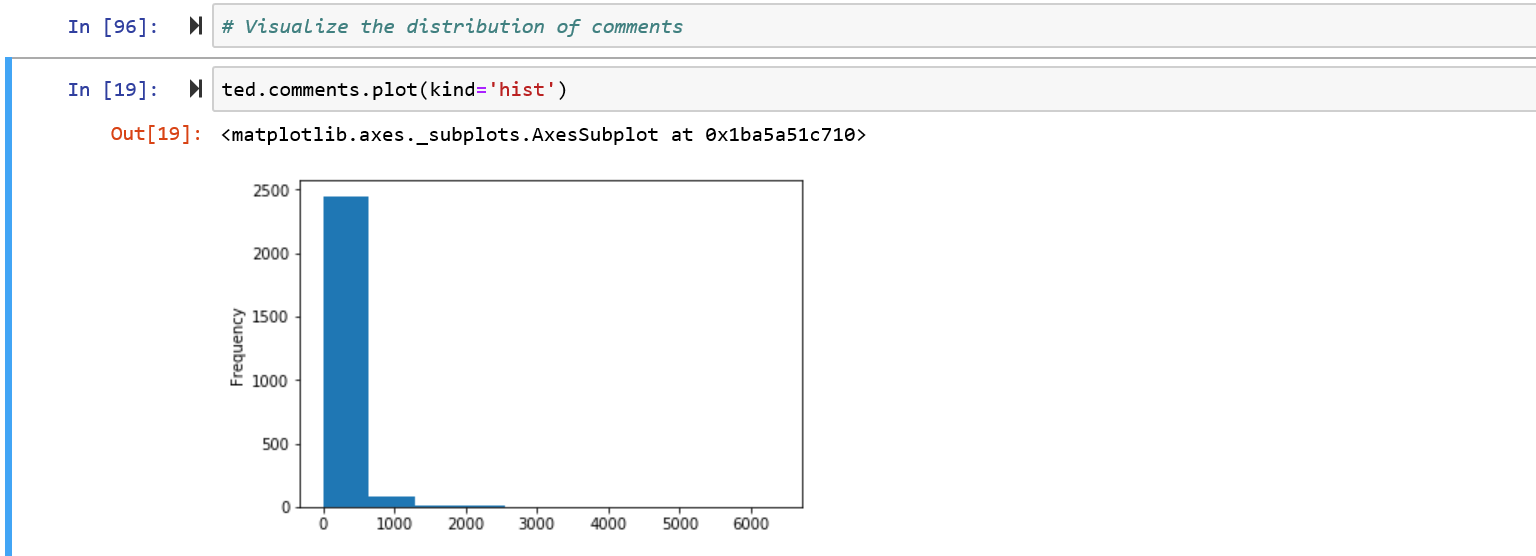
* 1. **Graph of film datetime vs year ( Without Sorting ).**

****

* 1. **Graph of film datetime vs year ( With Sorting ).**

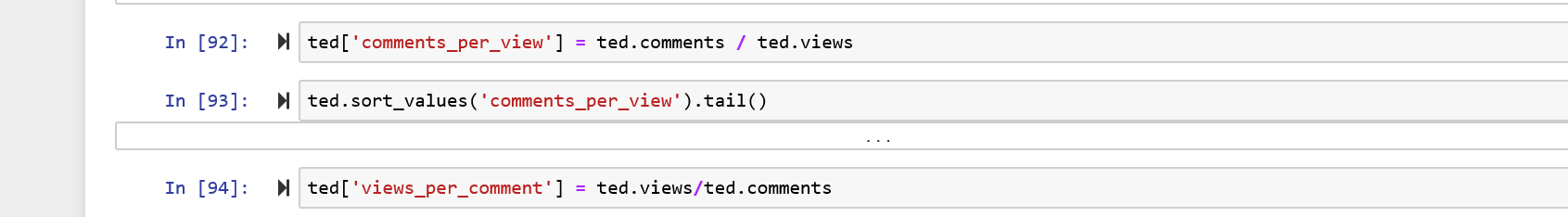
****

* 1. **Histogram plot of comments.**

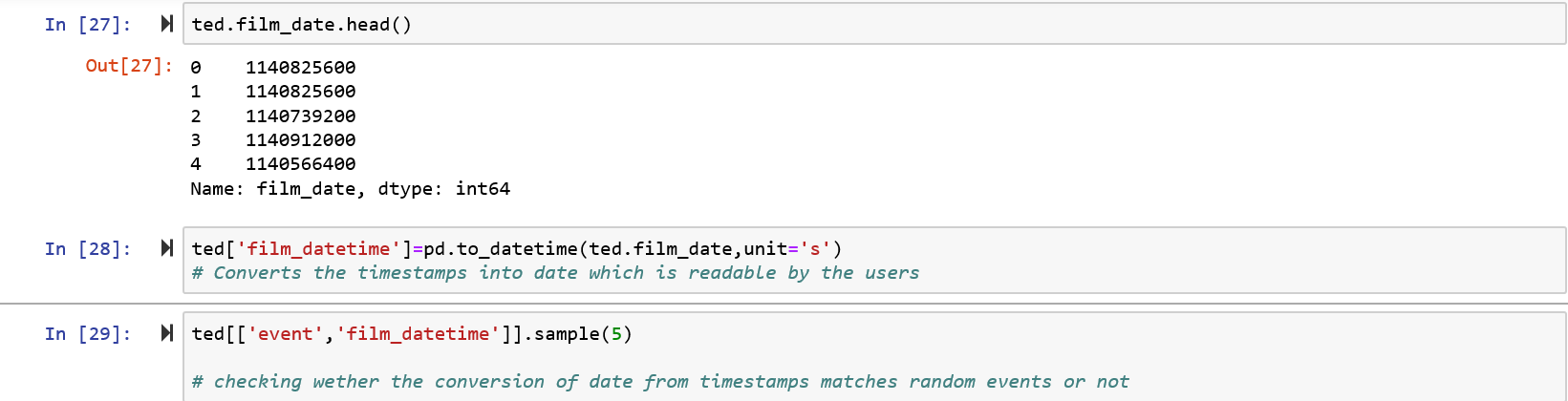
****

**Data Cleaning:**

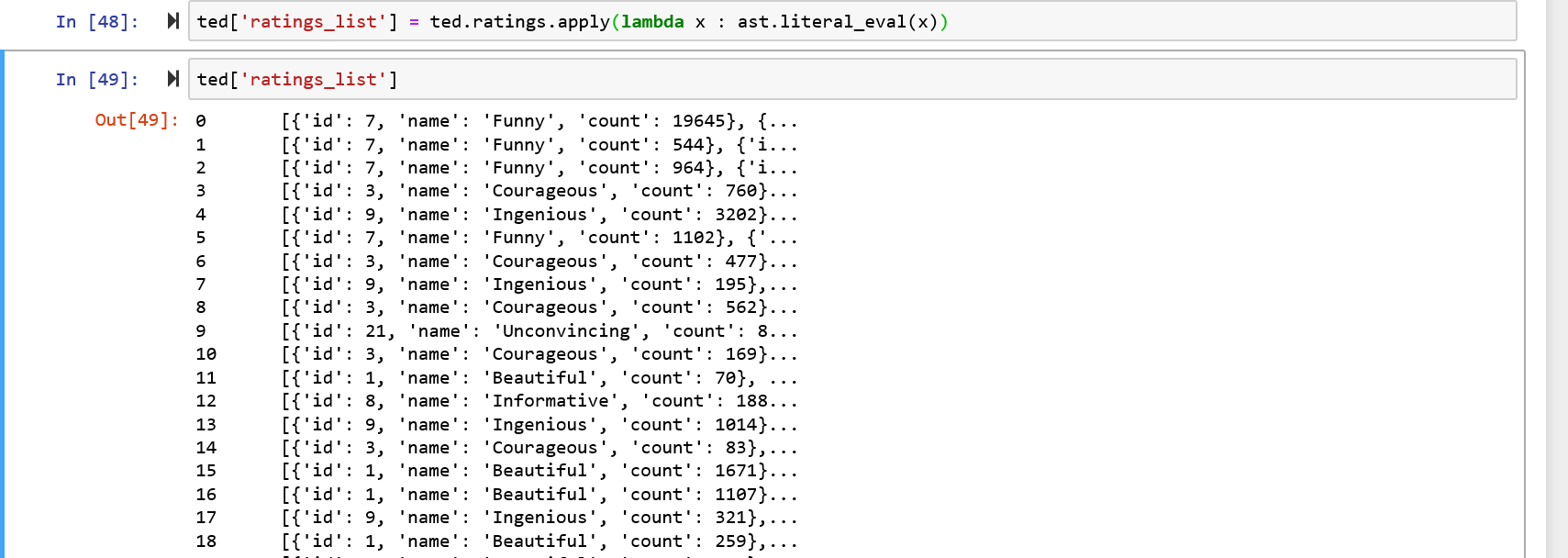
**1.We create separate columns named comments\_per\_view and views\_per\_comment.**

****

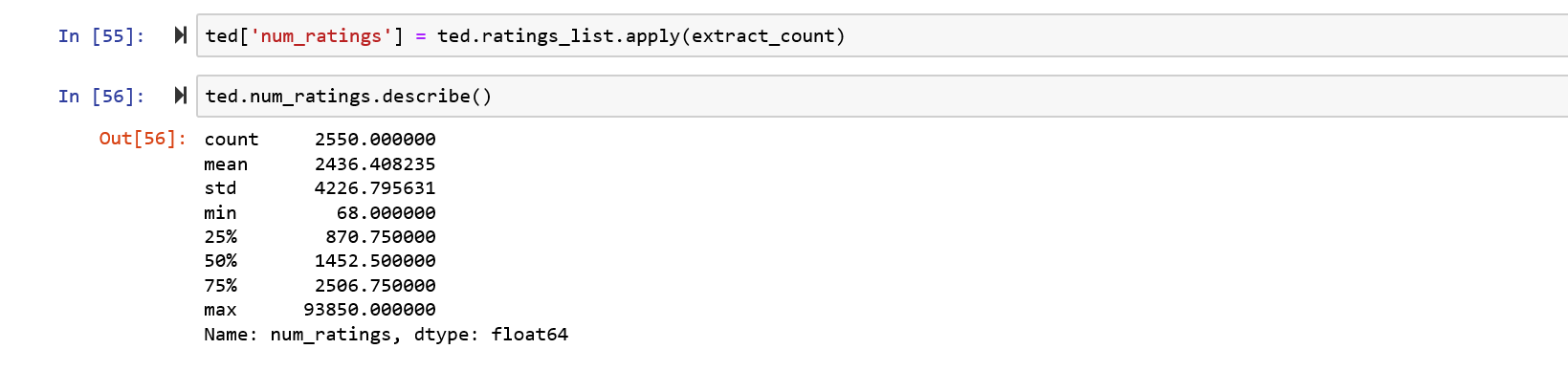
**2. We converted the timestamps into date which is readable by the users.**

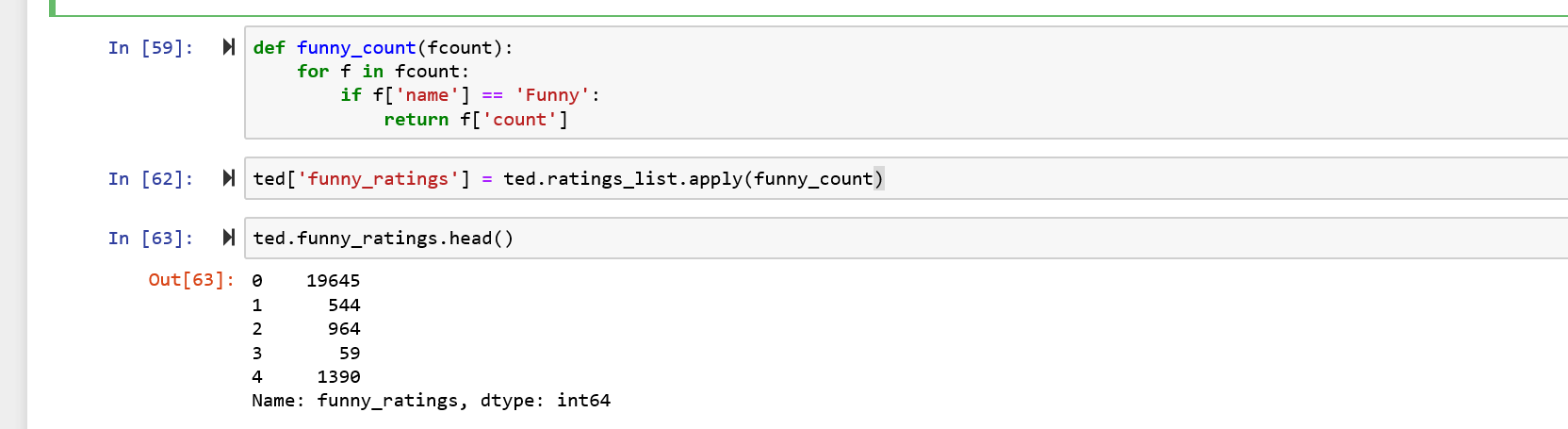
****

**3. We also introduced a new column named ratings\_list to cope up with the solution for unpacking the ratings data.**

****

1. **We also added new columns named num\_ratings and funny\_ratings to find the total number of ratings received by each talk and count the number of funny ratings respectively.**

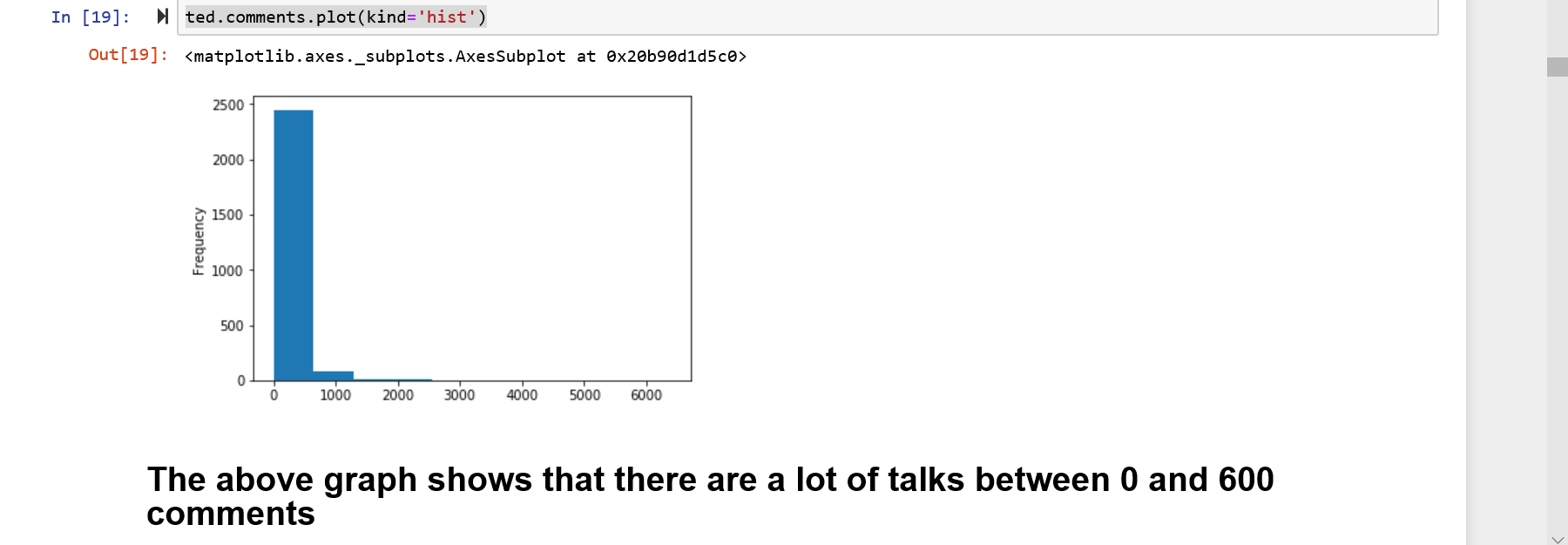
****

****

**Result and Discussion**:

**Result:**

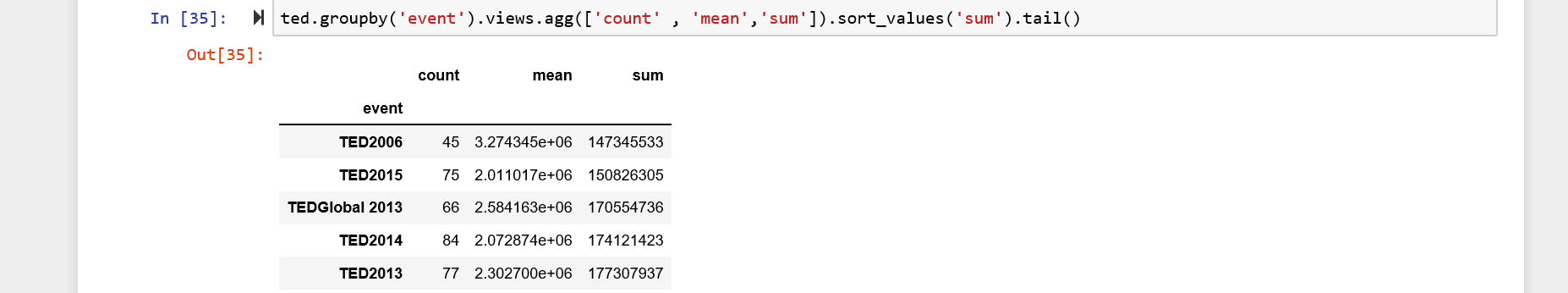
1. **Which talks provoke the most online discussion ?**

****

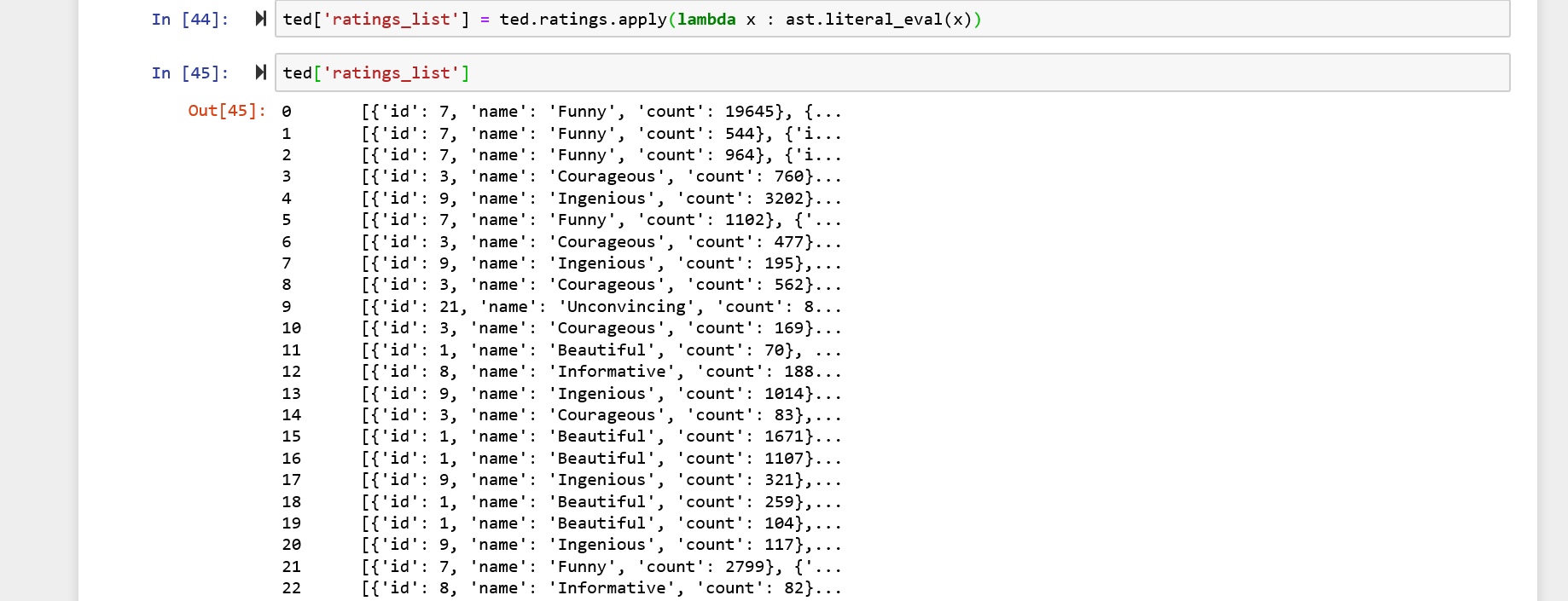
1. **Plot the number of talks which took place each year .**

****

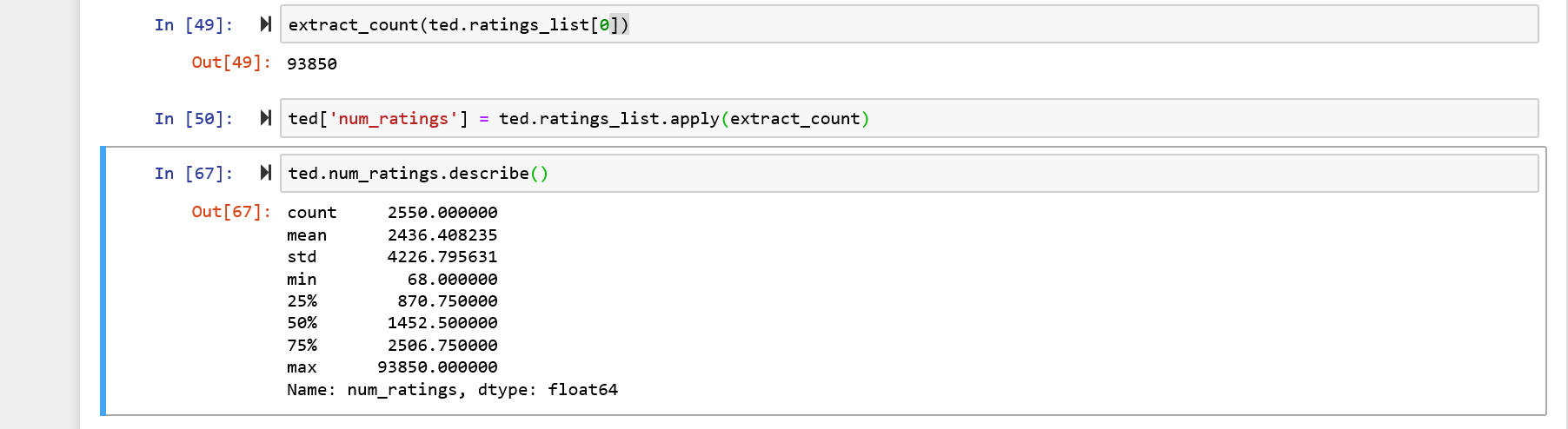
1. **What were the best events in Ted history to attend ?**

****

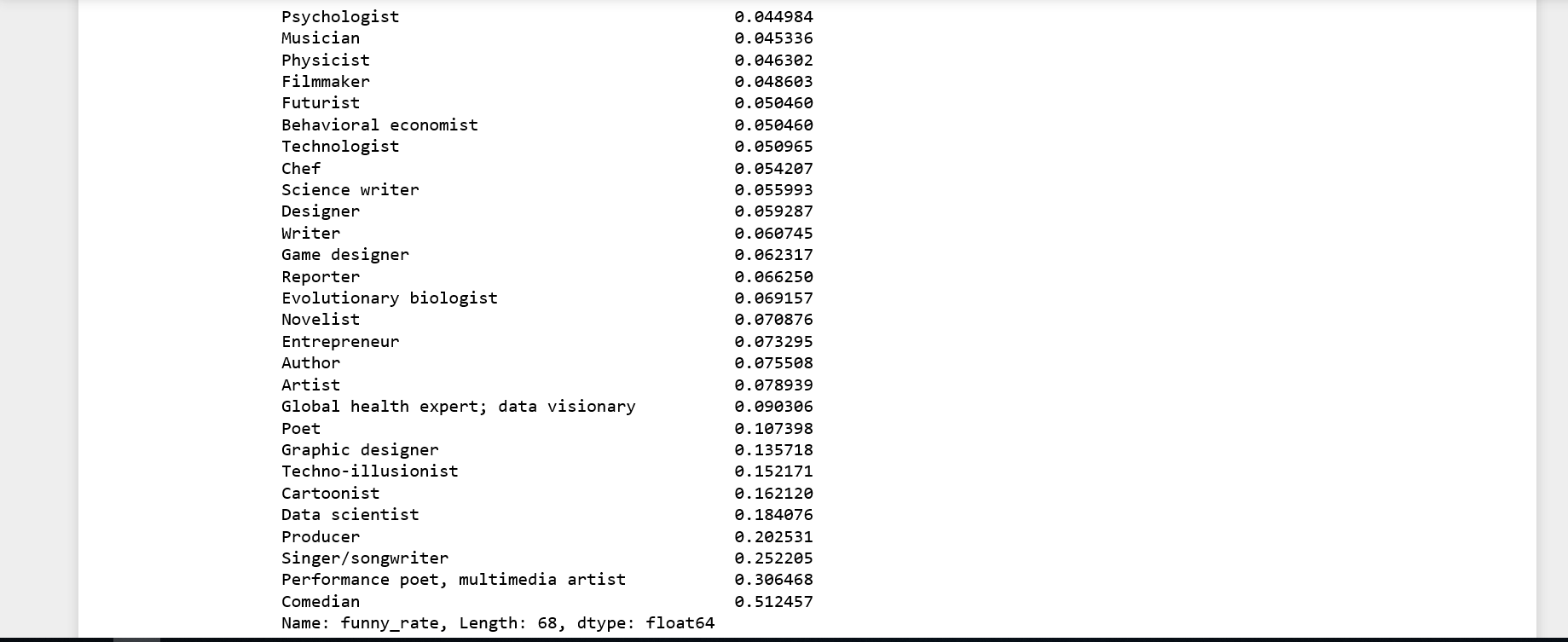
1. **Unpack the ratings data.**

****

1. **Count the total number of ratings received by each talk.**

****

1. **Which occupations deliver the funniest ted talks on an average ?**

****

**Discussion:**

The dataset here is a history of Ted Talks delivered by different speakers on a particular date . It also has the ratings for each talk and the comments as well .The user wants to know about each talk in different perspective like views , coments , popularity , etc . Specifically, here the problem is to modify the data table in a creative way to get in depth analysis of this Ted Talks Dataset. In this dataset we have 2550 rows and 17 columns. The analysis part includes count and histogram plot of comments and ratings . We also added few columns namely num\_ratings , ratings\_list ,etc to get the favourable results in an efficient way .

**Conclusion:**

1. The talks which have comments count between 0 and 600 provoke the most online discussion.

2. The highest number of talks took place in the year 2013 ( 270 Talks ) and in the year 1972 only 1 Talk took place.

3. The best events to attend in TED History are TED2013 and TED2014.

4. 93,850 is the highest number of ratings received among 2,550 talks.

5. Occupations like data scientist , producer , singer and comedian deliver the funniest Ted Talks on an average.

**References:**

https://www.python.org/

https://anaconda.org/anaconda/python

http://www.numpy.org/

https://matplotlib.org/

http://scikit-learn.org/

https://pandas.pydata.org/

https://pandas.pydata.org/

https://ipython.org/