Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Homework
- Basic stats
- HW for Weds

HW7

- Questions that you wish to ask of your data set. Specifically write:
    - What is the question?
    - What data will be used to answer the question?
    - Describe a second way to answer the same question?
    - What are assumptions and limitations of the approach?
    - What is the statement that you will be able to make after doing the analysis?

- The Dataset used is NYC traffic volume counts
- Questions to be asked during the analysis of this dataset.
    - 1. What does this data signify?
    - 2. Which direction is the busiest?
    - 3. Which time of the day is busiest?
    - 4. Which route needs more busses?
- Answers to the above questions.
    - 1. The quickest route can be used by analyzing the volume during different time of the day.
    - 2. The answer to which routes needs more busses will be given by analyzing the traffic count and transportation count for that particular route.
- By the analysis of this dataset we can find the best route to choose during a particular time of the day and also which routes to completely avoid during a particular time.

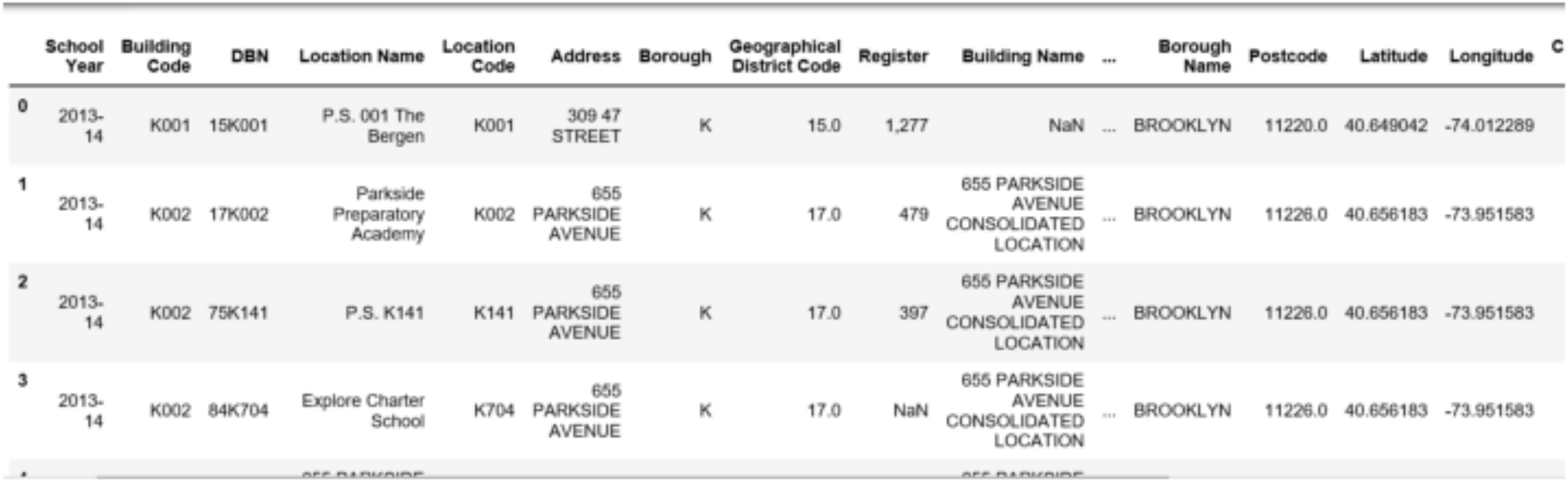
HW7

- Questions that you wish to ask of your data set. Specifically write:
    - What is the question?
    - What data will be used to answer the question?
    - Describe a second way to answer the same question?
    - What are assumptions and limitations of the approach?
    - What is the statement that you will be able to make after doing the analysis?

<u>Questions to be asked on a data set.</u>

Here I am taking New York School Safety Report data set.

I did EDA on this data set. This data set is regarding safety report of the school location. The data set is loaded and displayed below.



	School Year	Building Code	DBN	Location Name	Location Code	Address	Borough	Geographical District Code	Register	Building Name	...	Borough Name	Postcode	Latitude	Longitude	C
0	2013-14	K001	15K001	P.S. 001 The Bergen	K001	309 47 STREET	K	15.0	1,277	NaN	...	BROOKLYN	11220.0	40.649042	-74.012289	
1	2013-14	K002	17K002	Parkside Preparatory Academy	K002	655 PARKSIDE AVENUE	K	17.0	479	655 PARKSIDE AVENUE CONSOLIDATED LOCATION	...	BROOKLYN	11226.0	40.656183	-73.951583	
2	2013-14	K002	75K141	P.S. K141	K141	655 PARKSIDE AVENUE	K	17.0	397	655 PARKSIDE AVENUE CONSOLIDATED LOCATION	...	BROOKLYN	11226.0	40.656183	-73.951583	
3	2013-14	K002	84K704	Explore Charter School	K704	655 PARKSIDE AVENUE	K	17.0	NaN	655 PARKSIDE AVENUE CONSOLIDATED LOCATION	...	BROOKLYN	11226.0	40.656183	-73.951583	

Questions:

1) Which is the safest borough in which a school is located?
2) How many number of schools are situated in the safest area?
3) Can we predict number of students taking admission in the schools which are based in the safest area?
4) Can we predict the number of overall crime held in a particular year.

Data used to answer the questions:

Questions:

1) Which is the safest borough in which a school is located?
2) How many number of schools are situated in the safest area?
3) Can we predict number of students taking admission in the schools which are based in the safest area?
4) Can we predict the number of overall crime held in a particular year.

Data used to answer the questions:

We need year, location name , borough name, Number of student registered, Number of major crimes, number of other crime, number of property crime, building name, building code.

Assumptions:

There may be missing values. There may be few reputed schools which does not want to report crime to protect their image.

There may be a school which is closed but still the data is present.

Analysis:

The area which has the highest number of crime is Manhattan.

The safe areas for school is Brooklyn and highest number of students register there.

HW7

- Questions that you wish to ask of your data set. Specifically write:
    - What is the question?
    - What data will be used to answer the question?
    - Describe a second way to answer the same question?
    - What are assumptions and limitations of the approach?
    - What is the statement that you will be able to make after doing the analysis?

1) What is the question?
ANS. What Genre games and on which Platform should the Publishers concentrate on growing more in the future in view of the user ratings, critic scores and their sales.

2) What data will be used to answer the question?
ANS. NA_Sales, EU_Sales, JP_Sales, Other_sales, Global_sales, Critic_score, User_score, User_count, will be used to predict what Genre games and on which platform the publishers should focus on developing in the Future.

3) Describe a second way to answer the same question?
ANS. We can get extra information for the same from various sources available online and perhaps add more parameters to test our analysis.

4) What are the assumptions and limitations of the approach?
ANS. For the given dataset, the assumption would be the User_Score, which is given by the subscribers of the game and the User_Count. Only one out of every odd client will have a similar rating for a specific game. One client may like it and give a 5 star review while the other client may very well give a 2-3 star review for a similar game. Thus they ought to have some different parameters to settle on a choice. For instance, they could incorporate another parameter 'No_of_downloads' for a specific game .

5) What is the statement you will be able to make after doing the analysis.
ANS. Analysing the information of the sales and the user and critic scores, the Publishers can grow more games in the specific Genre and on the Platform that is loved greatest by the users. Doing this the Publishers will have the capacity to raise their sales.

Dataset link: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

Basic statistics

- Chi square
- t-test
- correlation
- regression

Basic question:

- Are X and Y related?
  - HARD!
  - How to decide related vs not?
  - What kind of variables?
  - What can we assume to be true?

Neyman-Pearson statistical inference

- Measure how far your sample is from what one would expect based on random chance

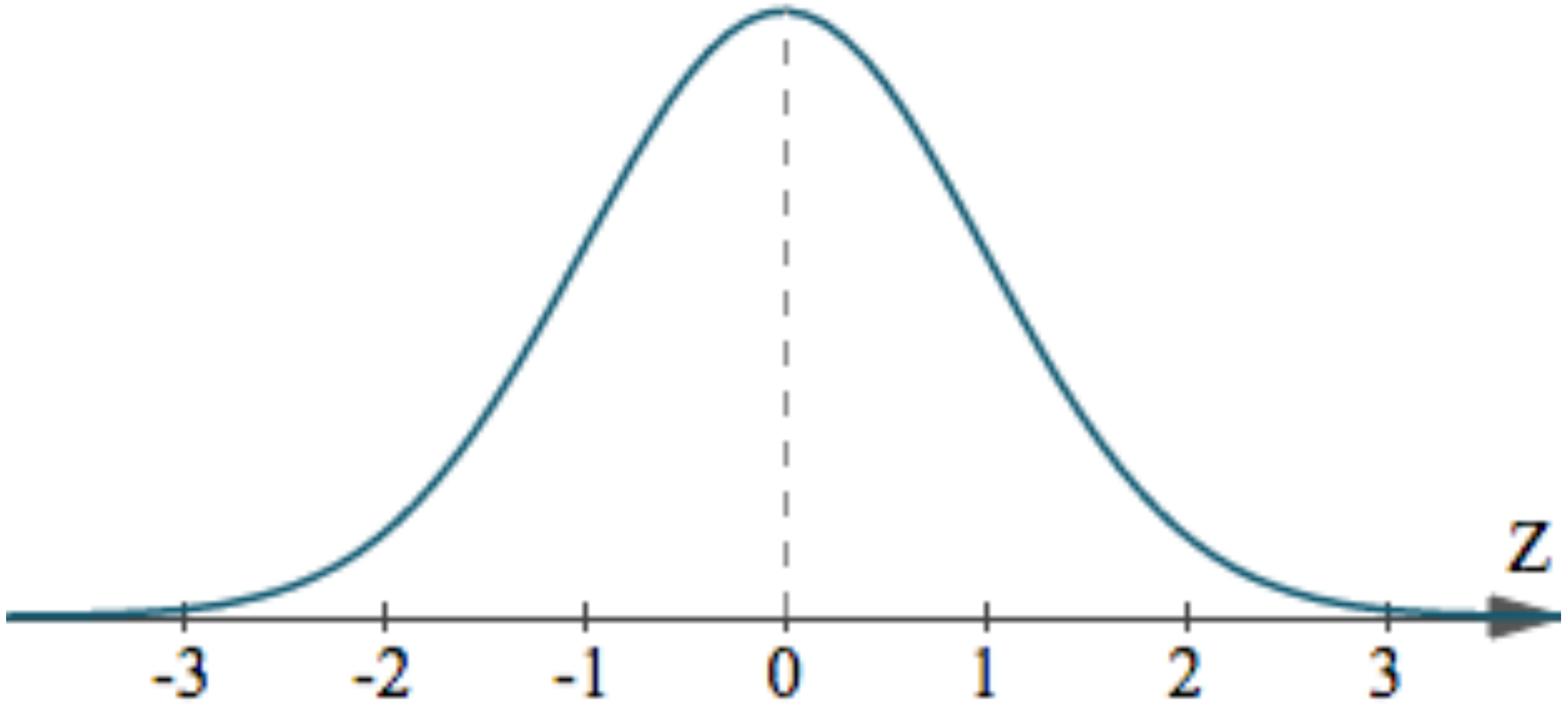
- Decide on a cut-off that is “far enough” (alpha)

    - This is typically 0.05. Meaning that there is a 0.05 chance or less of this event or more extreme assuming random sampling

Neyman-Pearson statistical inference

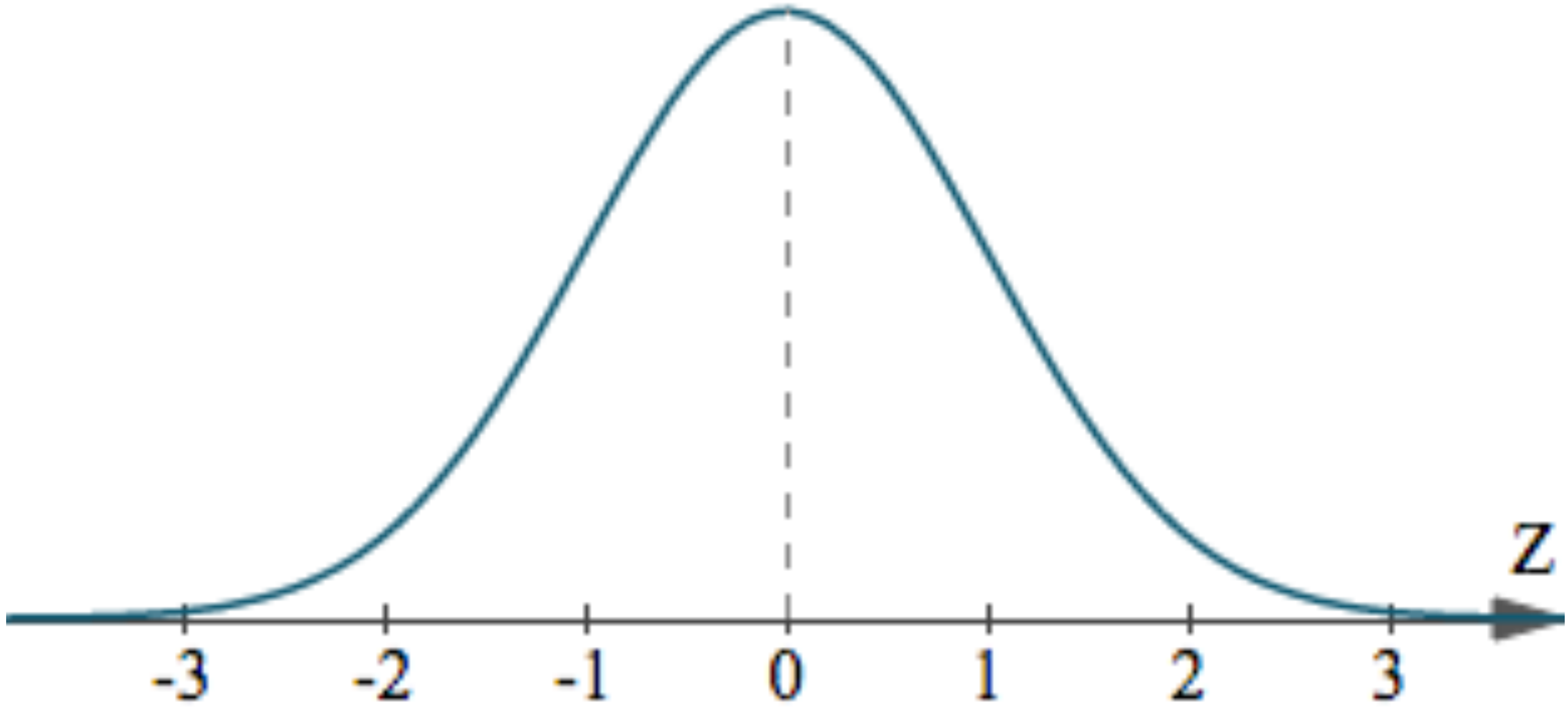
- Comparing two hypotheses:
    - Null hypothesis: There is no relationship/difference between the groups
    - Alternative hypothesis: There is a relationship/difference
- Neyman-Pearson only tests the first hypothesis!

Normal distribution



Assume we are talking about IQ scores.
The mean is 100, standard deviation is 15.
Thus, -1 on this plot corresponds to 85.

The region beyond which .05 of the mass resides is =/- 1.96.



Say we do a study in which we aim to make people smarter, and the average of our group is had a (standardized score) of 3. What can we say?

If the average had a standardized score of 1.
What could we say?

Statistical decisions: Knowing the possibilities

Truth

Decision

	Null hypothesis is **true**	Null hypothesis is **false**
reject Null hypothesis		
accept Null hypothesis		

Statistical decisions: Knowing the possibilities

Truth

		Null hypothesis is **true**	Null hypothesis is **false**
Decision	**reject** Null hypothesis	Type 1 error	
	accept Null hypothesis		Type 2 error

Statistical decisions: Knowing the possibilities

		Truth	
		Null hypothesis is **true**	Null hypothesis is **false**
Decision	**reject** Null hypothesis	Type 1 error (alpha=.05)	Power
	accept Null hypothesis	1-alpha	Type 2 error

HW8

- Improve your answers to HW 7. Make sure you answer every question. Be as precise as possible. Provide enough information for your answer to be evaluated.