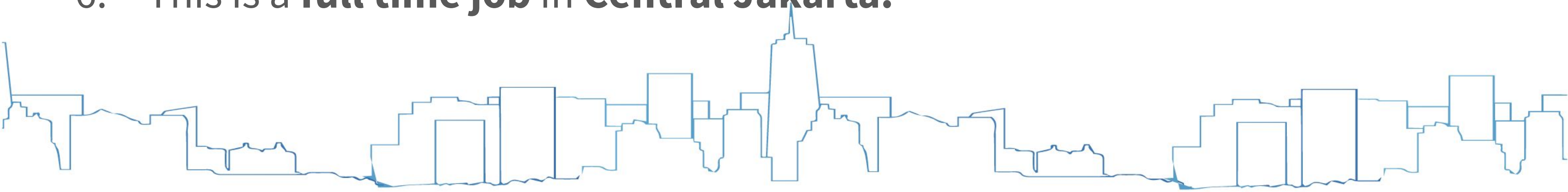
Data Science Interview Preparation

Deadline	27 April 2020
Interview Call	4-8 May 2020
Announcement	13 May 2020

(Read Clue Carefully in next slide)

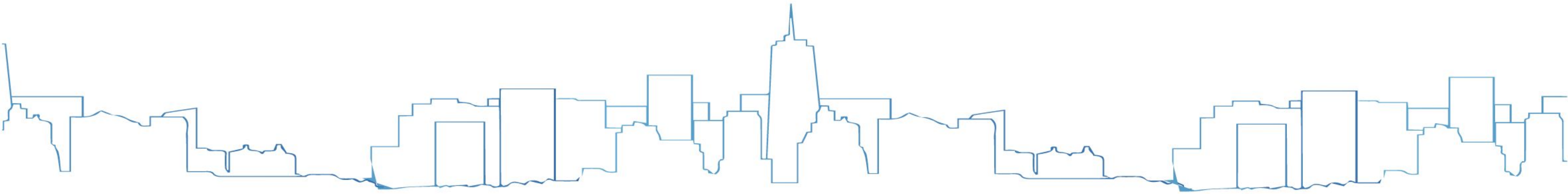
Instruction

1. In the interview process you are asked to answer your homework first in pdf format (It is recommended to convert from presentation software such as MS Powerpoint or Google Slides) no later than **April 27, 2020**.
2. Screenshots from outside are allowed as long as **the source is listed**.
3. Please solve the problem effectively, **minimize work** on create synthetic data, code, visualization, etc. Bring simplest answer that you can defend to technical and non-technical people effectively.
4. You can answer as soon as possible, and we can invite you for an **early interview.**
5. All interview processes are conducted **online**.
6. This is a **full time job** in **Central Jakarta.**

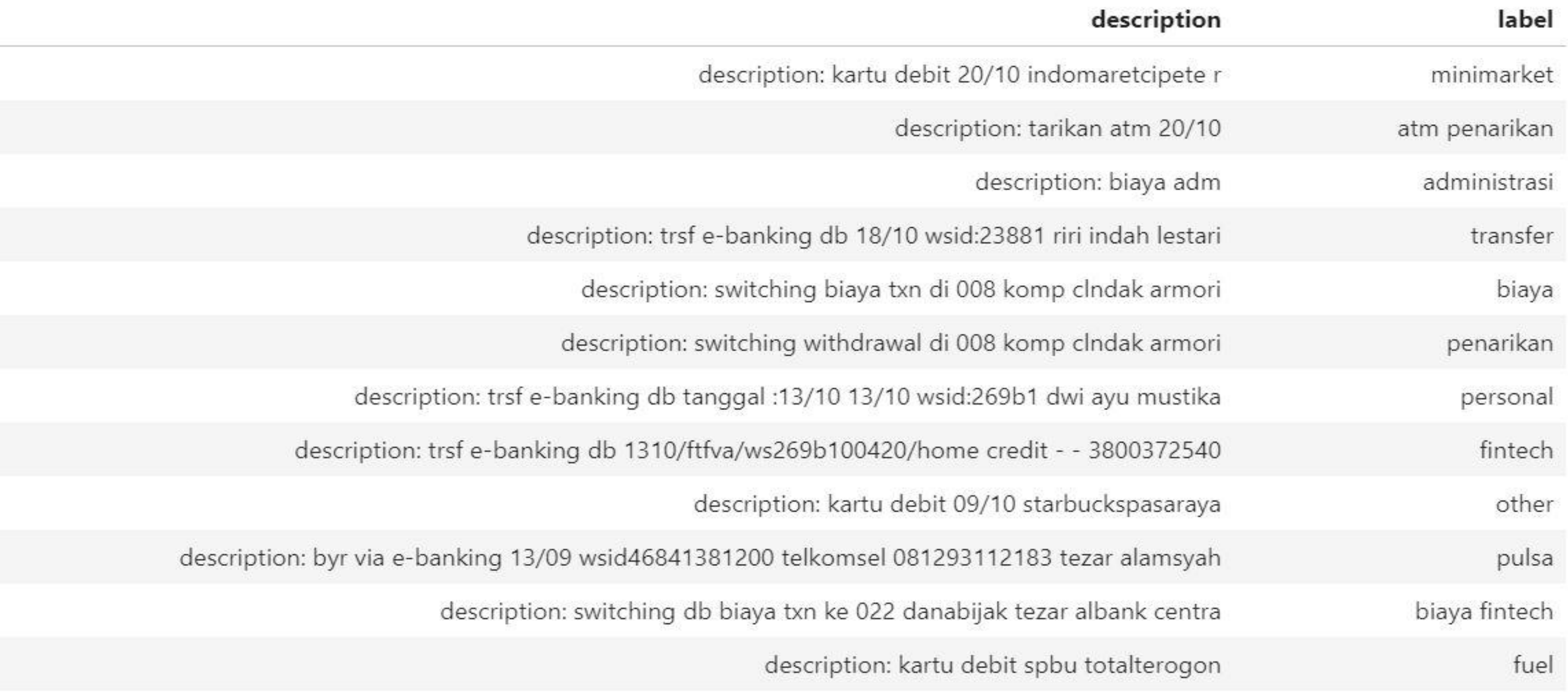


Materials

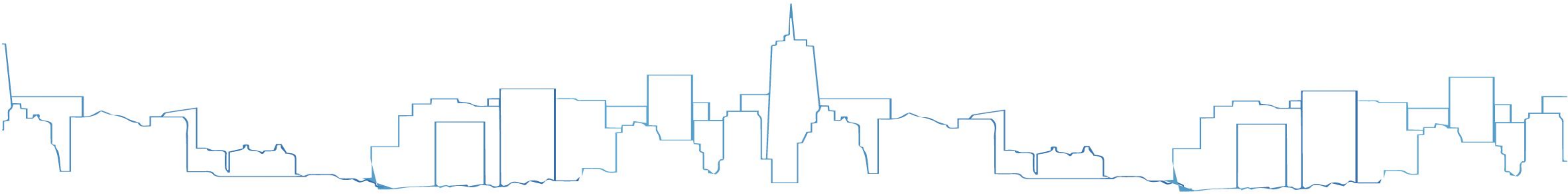
1. Data Cleansing
2. Data Analysis
3. Data Storytelling
4. Structured Thinking
5. Computer Vision
6. NLP
7. Statistics
8. SQL



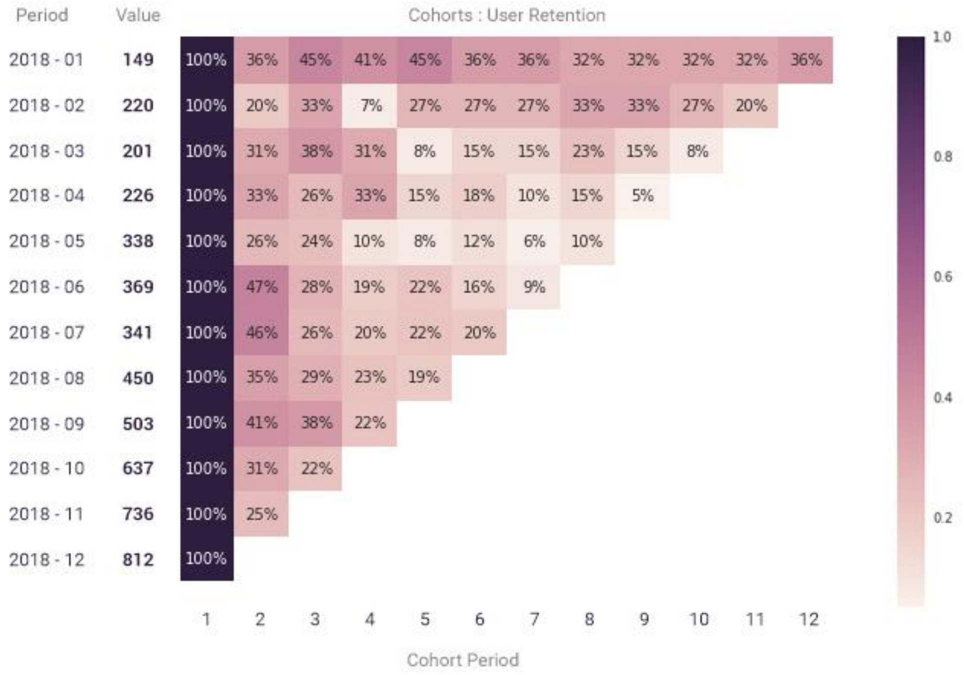
Dataset 1



description	label
description: kartu debit 20/10 indomaretcipete r	minimarket
description: tarikan atm 20/10	atm penarikan
description: biaya adm	administrasi
description: trsf e-banking db 18/10 wsid:23881 riri indah lestari	transfer
description: switching biaya txn di 008 komp clndak armori	biaya
description: switching withdrawal di 008 komp clndak armori	penarikan
description: trsf e-banking db tanggal :13/10 13/10 wsid:269b1 dwi ayu mustika	personal
description: trsf e-banking db 1310/ftfva/ws269b100420/home credit - - 3800372540	fintech
description: kartu debit 09/10 starbuckspasaraya	other
description: byr via e-banking 13/09 wsid46841381200 telkomsel 081293112183 tezar alamsyah	pulsa
description: switching db biaya txn ke 022 danabijak tezar albank centra	biaya fintech
description: kartu debit spbu totalterogon	fuel



Dataset 2

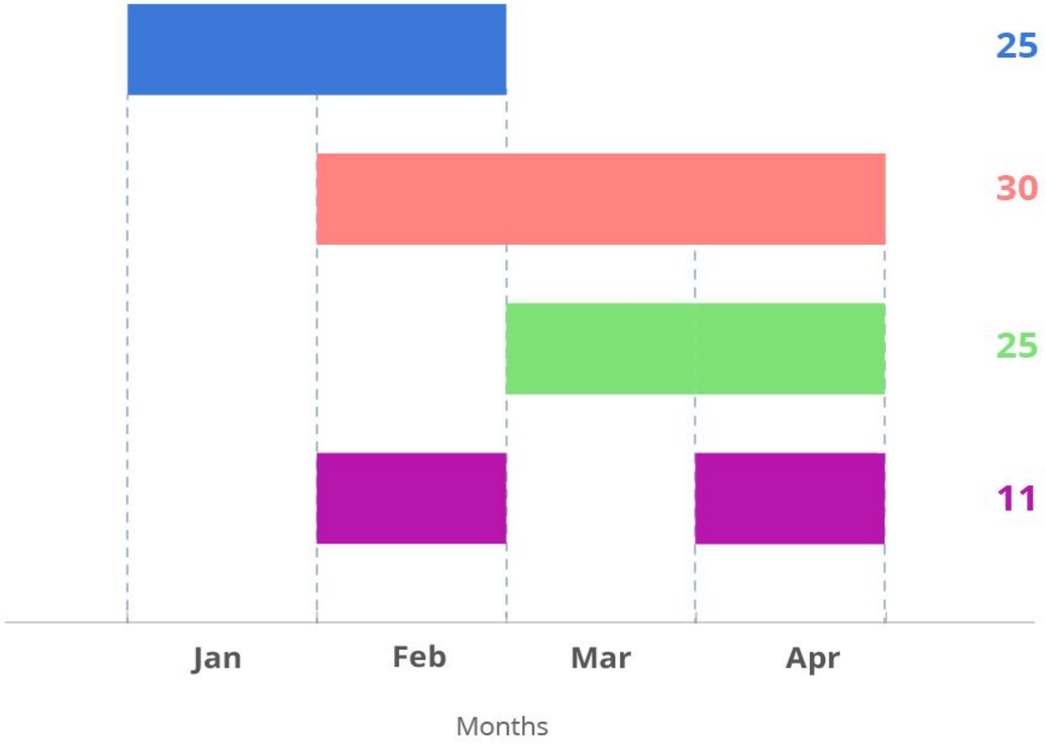


Cohorts : User Retention

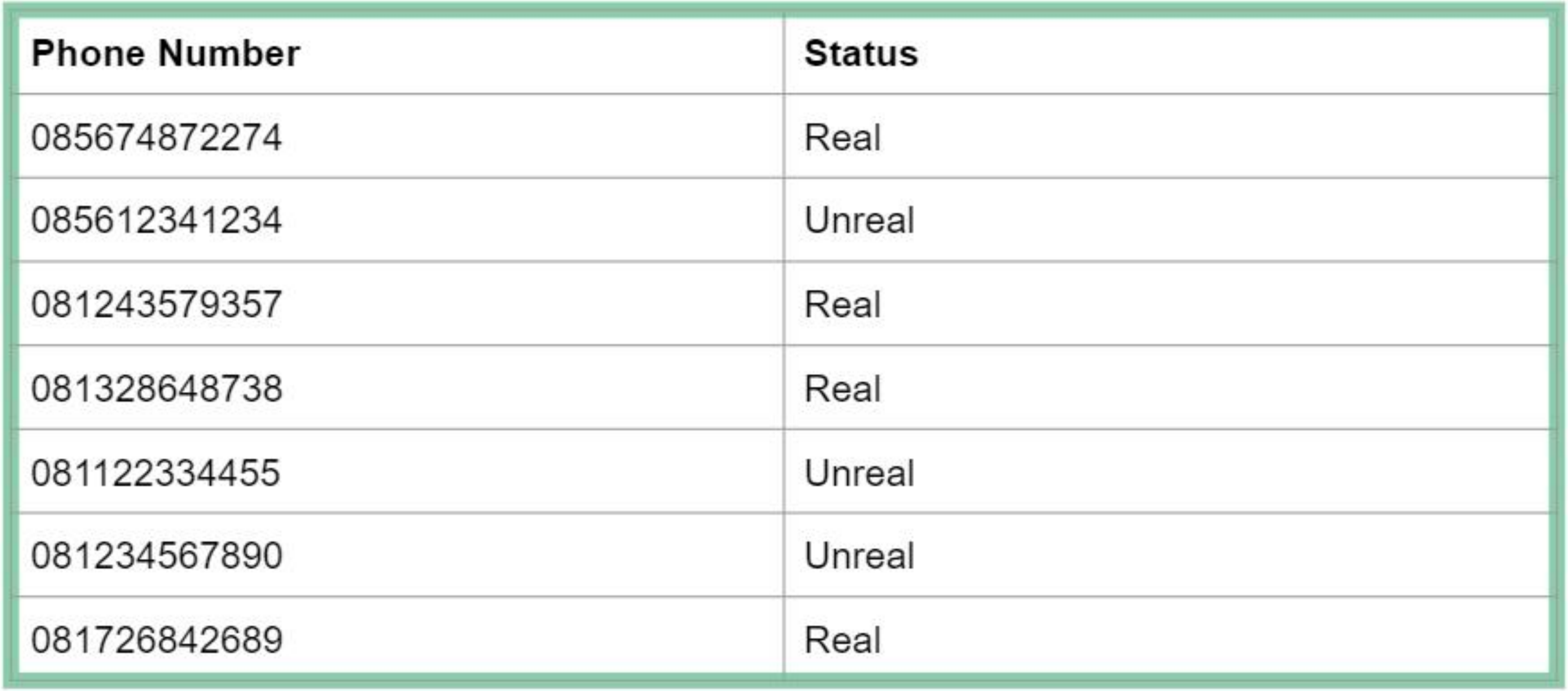
Period	Value	1	2	3	4	5	6	7	8	9	10	11	12
2018 - 01	**149**	100%	36%	45%	41%	45%	36%	36%	32%	32%	32%	32%	36%
2018 - 02	**220**	100%	20%	33%	7%	27%	27%	27%	33%	33%	27%	20%	
2018 - 03	**201**	100%	31%	38%	31%	8%	15%	15%	23%	15%	8%		
2018 - 04	**226**	100%	33%	26%	33%	15%	18%	10%	15%	5%			
2018 - 05	**338**	100%	26%	24%	10%	8%	12%	6%	10%				
2018 - 06	**369**	100%	47%	28%	19%	22%	16%	9%					
2018 - 07	**341**	100%	46%	26%	20%	22%	20%						
2018 - 08	**450**	100%	35%	29%	23%	19%							
2018 - 09	**503**	100%	41%	38%	22%								
2018 - 10	**637**	100%	31%	22%									
2018 - 11	**736**	100%	25%										
2018 - 12	**812**	100%											

Cohort Period

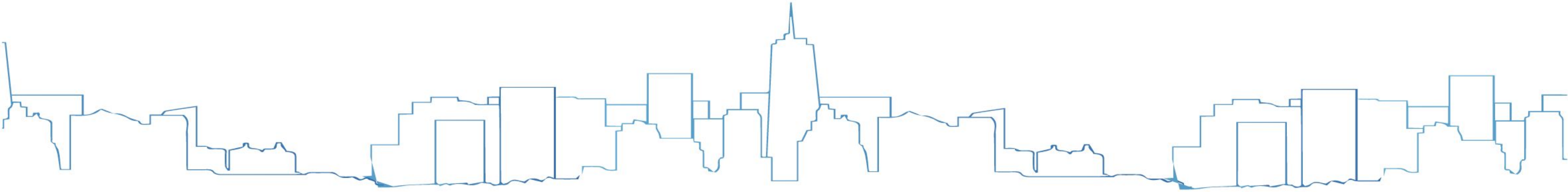
Dataset 3



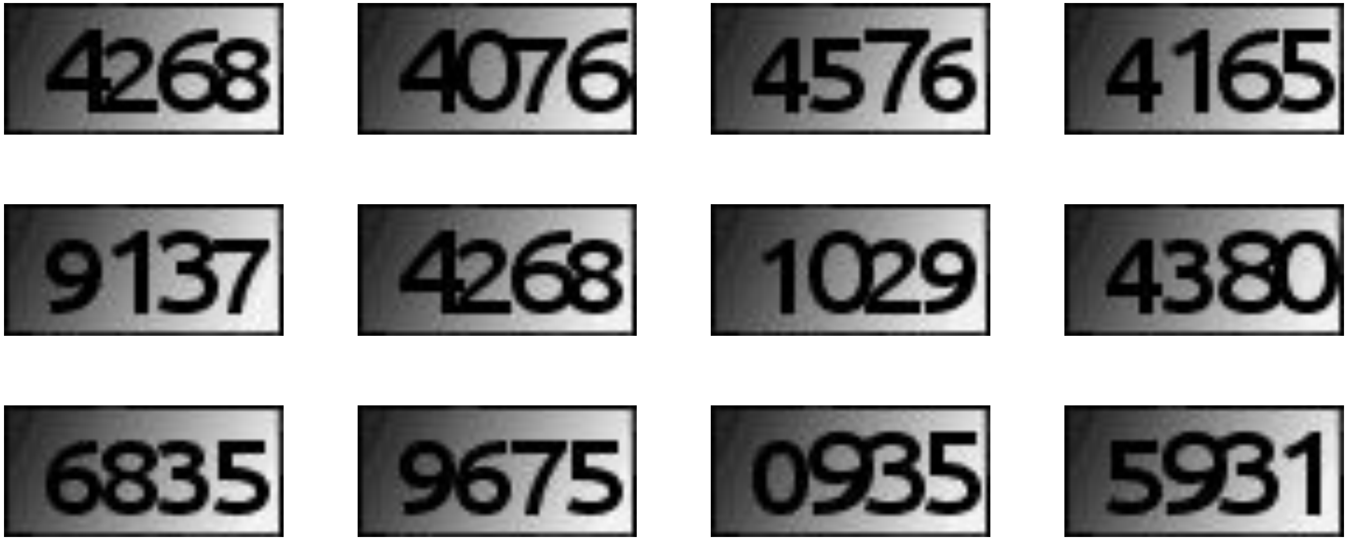
Dataset 4

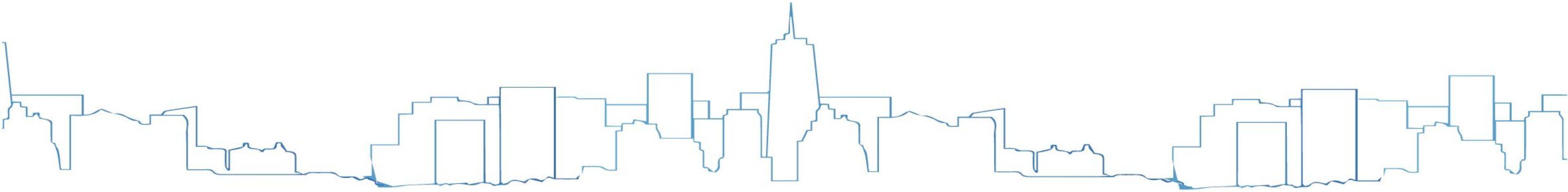


Phone Number	Status
085674872274	Real
085612341234	Unreal
081243579357	Real
081328648738	Real
081122334455	Unreal
081234567890	Unreal
081726842689	Real



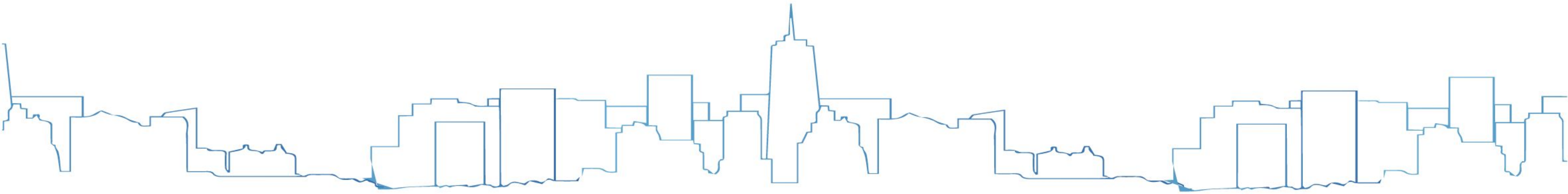
Dataset 5





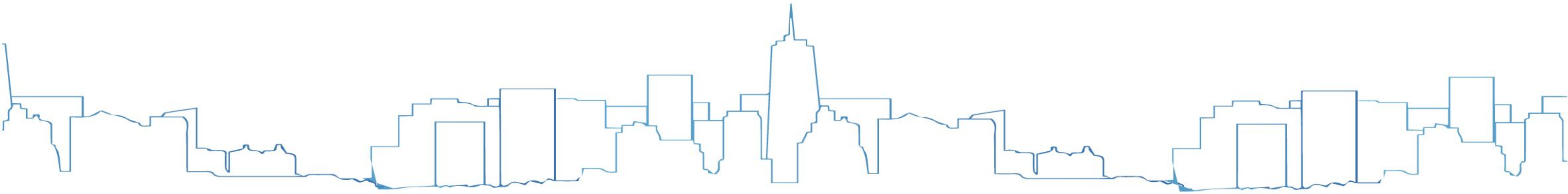
Problem 1

1. In Dataset 1, How to transform the description column in order to make it easier to analyze?
2. Assume that you have 15 millions rows of data. If the columns `label` is empty in 10 millions rows what will you do to fill the missing data?
3. What yo do to deal with abbreviation and misspelled words?
4. How to deal with Imbalanced Classes, Outliers, and Rare Data?



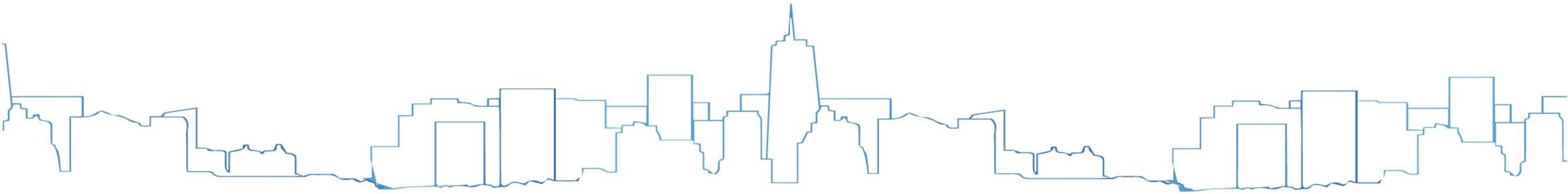
Problem 2

1. In Machine Learning, What is difference between bias and variance?
2. How do you know if one machine learning algorithm is better than another on accuracy, reliability, and scalability?
3. What is difference between close-form and non close-form model in Machine Learning?
4. What is difference between feature, parameter, and variables?
5. What is differences between Hold-Out Validation, Cross-Validation, and Bootstrapping?



Problem 3

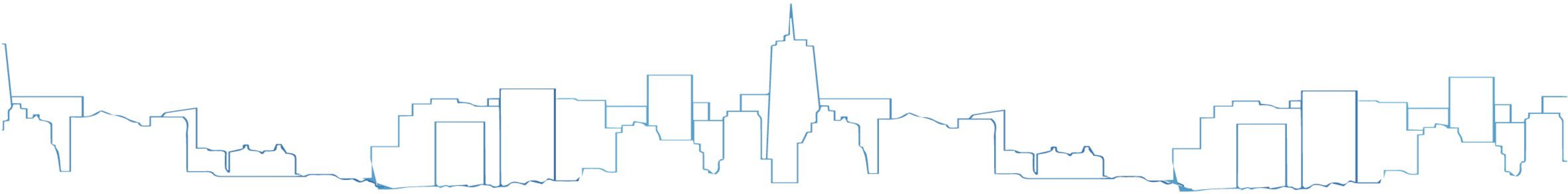
1. Based on Dataset 2, how many people that came in May 2018 are still coming in July 2018?
2. What data is needed to build chart on Dataset 2 and Dataset 3?
3. Write pseudocode for create chart on Dataset 2!
4. Write pseudocode for create chart on Dataset 3!
5. If you make chart on Dataset 3, what chart that you need to make?



Problem 4

1. Based on dataset 4, What pattern determined that the number is real and unreal?
2. Write pseudocode to determine if the number is real and unreal?

(Clue: you can do multiple pseudocode)

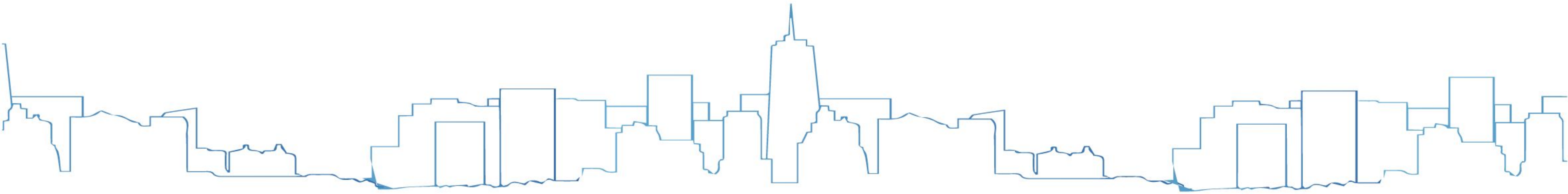


Problem 5

1. Describe the required steps in order to build a proper captcha reader engine (assume you use Dataset 5)
2. What is the difference between Semantic Segmentation, Object Detection, Image Generation, and Pose Estimation in terms of Input, Output and Label?
3. In YOLO (https://pjreddie.com/darknet/yolo/), there are 5 type of loss function, can you please explain them?

Problem 6

1. Explain differences (pros and cons) between building chatbot with NLTK, Seq2seq, and Rasa Framework
2. What is differences between TF-IDF, Cosine Similarity, FastText in terms on text based feature engineering?



Problem 7

Please fill this table

Problem	Input	Output
Clustering		
Survival		
Time Series		
Classification		
Recommendation Enggine		

Problem 8

1. What is difference between primary key, foreign key, and unique key?
2. What is difference between right, inner, full, left, and right join?
3. What is difference between unique, clustered and nonclustered index in SQL?
4. What is the difference between DELETE and TRUNCATE commands?
5. What is the command used to fetch first 5 characters of the string?

Thank you