Datanest

# Data Science Internship Interview Preparation

## Deadline: 1 July 2019

*(Please answer using the slide format and give your full name on the first slide on your answer slide)*

# Notes

Deadline: 1 July 2019

As fast as you apply as fast as you can be interviewed, but please prepare your answer carefully

# Title of Content

1. Data Cleansing
2. Data Analysis
3. Data Storytelling
4. Structured Thinking
5. Data Solution Implementation
6. Optional Research Capacity : Computer Vision
7. Optional Research Capacity : NLP
8. Optional Research Capacity : Bayesian Statistics
9. Optional Research Capacity : Frequentist Statistics
10. Optional Research Capacity : Feature Engineering

# Clue

In Datanest, we believe problem solving is the key of any data science activities.

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.

The problem is not hard, but it requires you to be resourceful and have a strong understanding of the problem

We start from “**Dataset 2**” so “**Dataset 1**” does not exist in this kit

# Optional

There’s 5 research capacities that we are assessing:

1. Research Capacity : Computer Vision
2. Research Capacity : NLP
3. Research Capacity : Bayesian Statistics
4. Research Capacity : Frequentist Statistics
5. Research Capacity : Featuring Accuracy

Your score is determined by 3 topics with the highest score, you can work on 5 of them .

# Dataset 2

| | 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 |

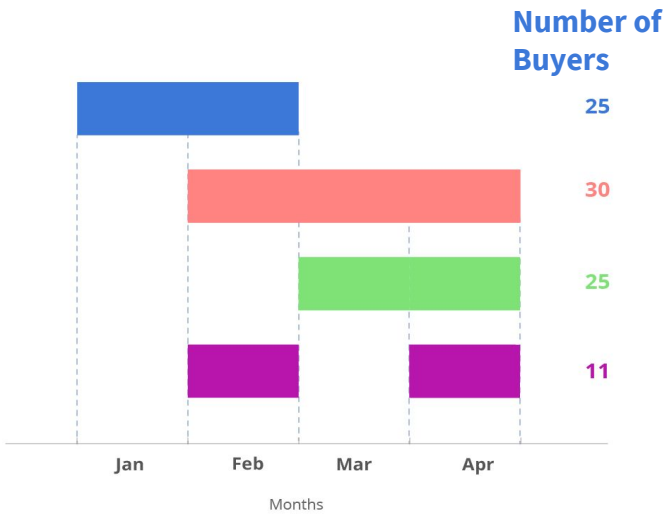
# Dataset 3

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% | | | | | | | | | | | |

1.0
0.8
0.6
0.4
0.2

Cohort Period



# Dataset 4

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

## Problem 1: Data Cleansing

1. In Dataset 2, How to transform the description column in order to make it easier to analyze?
2. 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: Data Analysis

1. 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?
4. What is difference between feature, parameter, and variables?
5. What is difference between survival analysis, time series analysis, classification, recommendation engine and clustering (in terms of input and output)?
6. What is differences between Hold-Out Validation, Cross- Validation, and Bootstrapping?

## Problem 3: Data Storytelling

1. Based on Dataset 3 (Slide 7) left chart, how many people that came in May 2018 are still coming in July 2018?
2. What data need to make chart on Dataset 3?
3. How to create the left chart on Dataset 3?
4. How to create the right chart on Dataset 3?
5. If we make chart based on left chart in Data, what chart that you need to make?

## Problem 4: Structured Thinking

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: Data Solution Implementation

1. What is differences between Business Intelligence  and Data Science in terms of (a) business question, (b) analytic characteristics, (c) analytic engagement processes, (d) data models, (e) business view.
2. List benefits that data lake could bring to organizations existing data warehousing environment, business analysts and data scientists.
3. What are issues that are preventing companies migrates to cloud solutions?
4. List the cultural changes that organizations must address if they would like to become data driven, to leverage big data to its maximum business potential and what are the organization needs to address those challenges.
5. Select two of outward-facing BI dashboards that can be checked daily/weekly (one example for retail industry, one for financial industry), and what is the most important insight to be displayed?

# Problem 6: Computer Vision

1. Describe the required steps in order to build a proper object detection engine!
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 7: NLP

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 8: Bayesian Statistics

1. What is differences between Bayesian and Frequentist statistics?
2. What is types of bayesian statistics  are available on Ludwig (https://uber.github.io/ludwig/), and describe their  inputs and outputs of them?

# Problem 9: Frequentist Statistics

1. You have multiple ads to experiment in a  campaign, explain the steps of experiment using (a) A/B Testing, and (b) Multi-Armed Bandit.
2. What is difference between panel data analysis, longitudinal data analysis, multilevel statistical model, and structural equation modeling in terms of dataset requirement?

# Problem 10: Feature Engineering

1. What is differences between LabelCount, Target , NaN Encoding, Polynomial, Consolidation and Expansion Encoding
2. What is differences between standard (Z), MinMax, Root  and Log scaling
3. Please list  5 feature engineering on address data

## Closing

Great things happen to those who don't stop believing, **trying**, learning, and being grateful.