Dataset Hunting

**Initial Thoughts:**

Selecting data that is unique and different than commonly analyzed data & selecting one that is not misleading as well.

Some of these dataset providing websites require you to create an account and then access resource and sometimes also permission/ fees to use their datasets. One could subscribe to a website that features daily news with certain attributes and those could be fetched by RSS Feeds that is, via XML.

**Asking the right question?**

WHAT IS THE USE (CASE) OF THIS DATA???

WHAT CONTEXT OF DATA WILL BEST YIELD RESULTS FOR MY WORK???

Also look out for different combination while dataset selection. The best suited for beginners are by going in sequence. Each has its own challenges of dealing with.

1. Structured-&-Static
2. Structured-&-Dynamic
3. Unstructured-&-Static
4. Unstructured-&-Dynamic

Here is a list of shortlisted links for further lead.

**NOTE:** The following list mostly covers Structured-&-Static or Structured-&-Dynamic.

**1. Data for Cause:**

- Energy Use Globally

<https://datasource.kapsarc.org/explore/dataset/world-energy-use-1960-2012/?disjunctive.country>

2. **World Happiness Data:**

<https://www.kaggle.com/unsdsn/world-happiness>

**3. Fascinating Lego Sets Combinations:**

<https://www.kaggle.com/rtatman/lego-database>

**4. Want to know about Bit coin transactions every 10 minutes:**

<https://www.kaggle.com/bigquery/bitcoin-blockchain>

**5. Zomato: In which country, what cuisine should I try and where?:**

<https://www.kaggle.com/shrutimehta/zomato-restaurants-data>

**6. Deep Dark Web: Where are the drug dealers or weapons dealers?**

<https://www.kaggle.com/philipjames11/dark-net-marketplace-drug-data-agora-20142015>

**9. AWS Honey Pot Attack:**

<https://www.kaggle.com/casimian2000/aws-honeypot-attack-data>

**10. Anybody up for analyzing Bhagavad Gita?**

<https://www.kaggle.com/schcsaba/bhagavadgita/data>

**11. Most popular website in this world?**

<https://www.kaggle.com/bpali26/popular-websites-across-the-globe>

**12. Google play-store Apps:**

<https://thewebminer.com/android-google-play-apps>

**13. Fraudsters:**

<https://www.kaggle.com/rtatman/fraudulent-email-corpus>

**14. AWS honeypot:**

<https://www.kaggle.com/casimian2000/aws-honeypot-attack-data>

**15. IPL Data:**

<https://www.kaggle.com/manasgarg/ipl>

Research scope: Predicting the winner of the next season of IPL based on past data, Visualizations, Perspectives, etc.

**16. TRIVAGO of DATASETS:**

<https://data.opendatasoft.com/explore/?sort=modified&disjunctive.language&disjunctive.source_domain_title&disjunctive.theme&refine.language=en>

**17. Data breach information:**

<https://haveibeenpwned.com/PwnedWebsites#OnlinerSpambot>

<https://haveibeenpwned.com/API/v2>

**18. Dynamic Dataset Links:**

- <http://www.secrepo.com/>

- <http://www.webiron.com/abuse_feed/>

- <http://cybercrime-tracker.net/index.php>

- <https://ransomwaretracker.abuse.ch/feeds/>

- <https://ransomwaretracker.abuse.ch/tracker/>