# **Product Analyst Self-Study Links**

**Introduction**

This resource is intended for aspiring data scientists / data analysts looking to get started in the industry. It is an informal list of resources compiled based on feedback from data scientists at Facebook and the wider tech community.

**What is the role of a Data Analyst?**

* [2 minute intro](https://www.youtube.com/watch?v=5C9wQdz0yEE) video of FB’s VP of analytics Alex Schultz and others explaining what Data Analytics does at FB
* [Longer video of Alex Schultz explaining analytics and growth in much more detail](https://www.youtube.com/watch?v=URiIsrdplbo)
* <https://data36.com/data-science-for-business/>

**This resource focuses on building 6 skill areas, +1 optional area.**

1. Quantitative: Statistics & Probability
2. Research: Study Design
3. Technical: SQL
   1. Technical (Optional): Python
4. Business: Defining Business Metrics
5. Communication: Spreadsheets and Visualizations
6. Real world: Data Street Smarts
7. User Behavior

**These are the core competencies that will help a beginner become an effective data analyst. The material linked here is all free and covers them systematically.**

If you are already an expert in any area, you may skip it, though you might want to look at each area to be sure you already know it.

(Machine learning is not covered, as it isn’t required for beginners at data analytics.)

1. **Statistics & Probability**

Statistics and probability are the mathematical foundations of data analysis. Where do data analysts use them?

* Studying representative sample populations, and extrapolating conclusions to the general population.
* Measuring how a predictor variable (such as which phone users have, which version of the app they have, etc) impacts the outcome of a key performance metric.
* And much more.

For this section, try this full free course -

[Statistical Reasoning - Open and Free](https://oli.cmu.edu/jcourse/webui/guest/activity.do?context=729df49b0a0001dc61c437228919663c)

(Create a free account or enter without an account to see the course content)

**Core statistical concepts for population data that will be covered in the course:**

* Common distributions
* Random sampling
* Power, effect size and sample size
* Confidence intervals and p values, p value hacking
* Mix shift
* Regression modeling
* Bayes Theorem

**Additional material on statistics beyond the course**

* [Confounding](http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_Confounding-EM/BS704-EP713_Confounding-EM2.html)
* [Combinatorics](https://www.youtube.com/watch?v=64DcVdMLqhg&list=PLjRL3OOnLvS97zWj2Kk4TdXQDGanWpu4i)
* [Simpson paradox](https://en.wikipedia.org/wiki/Simpson's_paradox)

**Additional supporting resources**

* Simple, straightforward explanations of core concepts -<https://statisticsbyjim.com/>
* (best navigated with<https://statisticsbyjim.com/glossary/>)
* Practice statistics and probability in the form of quizzes with answer explanations - <https://brilliant.org/>

**2. Study design**

Analysis of data is based on carefully designed structures that ensure valid comparisons that can answer the research questions.

For example, analysts often do controlled experiments (A/B tests) in which users are randomly assigned to 2 or more different app versions, then their behavior is tracked over time and they are compared on a set outcome metric. Each step of how this study is constructed (selection of users, assignment to groups, followup period, definition of outcome metric) must be done in accordance with defined rules in order to ensure the conclusions are valid.

Without proper study design, someone can easily be misled by spurious correlations that stem from how they selected the users, assigned them, the duration of the followup, or how they measured the outcome.

This section identifies the core knowledge you need, and links to it across a range of sites from both the tech and medical sectors - the math and logic are transferrable across both applications.

* Various types of studies, their purposes, their limitations
  + <https://statisticsbyjim.com/basics/random-assignment-experiments/>
  + <https://statisticsbyjim.com/basics/observational-studies/>
* A/B Test methodology, purposes, and limitations
  + <https://data36.com/statistical-significance-in-ab-testing/>
  + <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083073/> (A/B tests are identical to clinical trials and all the same principles apply.
  + <https://exp-platform.com/2019WebConfABTutorial/> (Theory and examples of real-world A/B tests from Microsoft Bing)
* Types of bias, quantifying bias, overcoming bias
  + <https://data36.com/statistical-bias-types-explained/>
  + <https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/biases>
* Confounder analysis by stratification, regression modeling
  + <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017459/>

**3. SQL**

Most of the data analysts will study is stored in databases.

SQL is an abbreviation for structured query language. It is the main tool for the analysts that allows them to work with large volumes of data stored in databases, easily join separate data sets and apply complex logic while performing data transformation or aggregation.

This is a very powerful tool, this is why it is important to know how to use it, what is available there. It’s also important to understand what a database is database and how to design it.

Below are the main things for you to learn about SQL:

* [Interactive tutorial including core syntax, code examples for common tasks](https://mode.com/sql-tutorial/introduction-to-sql)
  + Additional resource: [SQL basics tutorial with interactive tasks examples](https://sqlzoo.net/)
* Database design for analytical tables
  + <https://panoply.io/data-warehouse-guide/the-difference-between-a-database-and-a-data-warehouse/>
  + <https://searchdatamanagement.techtarget.com/definition/denormalization>
* Transforming data for easier analysis
  + <https://panoply.io/data-warehouse-guide/3-ways-to-build-an-etl-process/>
  + <https://en.wikipedia.org/wiki/Extract,_transform,_load>
* Data Quality Checks and Data Profiling
  + <https://en.wikipedia.org/wiki/Data_cleansing>
  + <https://en.wikipedia.org/wiki/Data_profiling#How_data_profiling_is_conducted>

**3a. Optional: Python**

Python is not strictly required for a beginner data analyst, and you can go a long way with the knowledge of statistics, SQL, data structures and study design.

However, knowledge of even the basics of Python makes it easier to automate analytical processes and to make them reproducible from end to end. These aspects make it easier to automate and scale your analysis, thus greatly improving the impact of your work.

Many on-line resources offer courses and tutorials for Python. Here are a couple we found especially helpful:

* The [Python for Everybody](https://www.py4e.com/) free text book (and MOOCs)
* Collection of [helpful tricks for working with the pandas package](https://nbviewer.jupyter.org/github/justmarkham/pandas-videos/blob/master/top_25_pandas_tricks.ipynb) for dataframe manipulations

**4. Business Metrics**

Now that you know how to choose which study design is the best fit for your needs, have technical skills to get and transform data as needed, understand the statistics which can be applied, it’s time to learn how to choose a good metric to measure product/feature success or failure. Choosing good KPI (key performance indicator) is the most important thing in analyst job. Statistical and technical skills won’t matter if you choose the wrong KPI to make business decisions.

**Overview**

* [Intro to KPIs](https://www.youtube.com/watch?v=hh2aqJ9L0lI)
* [The difference between surrogate metrics vs terminal endpoints](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4554958/)

**Basic metrics for consumer apps**

* [Active users](https://www.youtube.com/watch?v=YqhyHtxhVxY)
* [Acquisition and Retention](https://www.youtube.com/watch?v=Pr-WYkdxe_Y)
* [Cohort Retention](https://www.youtube.com/watch?v=VvYV0V-kxr0)

**How to use metrics in an organization**

Once an organization has defined a KPI, they want to adhere to it as their guide for all strategy of the company. What does this mean?

* Tying all analyses to that kpi - when people do analyses, they need to state their findings in terms of impact on the KPI. So, if you added a new button to your app, don’t report how many clicks it got! Report how much it increased daily active users (or whatever KPI you’ve chosen).
* Opportunity sizing against a kpi - when you decide what to prioritize, choose those projects with the biggest potential to drive up our KPI.
* Making projections and goals for a kpi - team goals should be defined either directly through the KPI, or, through a lever with a known relationship to the KPI. For example, if your company metric is active users, you might have an acquisition team who’s KPI is daily new users, and a product team who’s KPI is 28 day retention of users. Together, these are the main determinants of your daily active users as we move forward in time.
* Drilling into unexplained changes in that kpi - if your KPI moves unexpectedly, you need to investigate why. It could be a data error, or it could be something important, whether good or bad.

**5. Spreadsheets and Visualizations**

Your work as data analysts includes communicating our findings to business stakeholders. While SQL and statistics might be enough for you to “convince yourself” of a certain hypothesis, for a broader audience you will usually want to simplify the presentation of the data to enable them to quickly understand it and make a decision based on it.

Spreadsheets (for tabular presentation) and visualisations (for visual presentation) are the main 2 tools we’ll use for this purpose.

In addition to helping us communicate with others, these are also extremely helpful for us ourselves, in quickly identifying distributions, time series trends, and other patterns in a large data set.

Spreadsheets

* [A concise lesson in how to use spreadsheets well](https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989)

Visualizations

* [45 Minute video summary](https://www.youtube.com/watch?v=xPdrU3vyXuk)
* [How Humans See Data](https://www.youtube.com/watch?v=fSgEeI2Xpdc#action=share)

Additional resources about visualizations:

* Good practice e[xamples to work on](http://perceptualedge.com/examples.php)
* [Comprehensive book on visualizations](https://serialmentor.com/dataviz/index.html)

**6. Data Street Smarts**

In real life, your results will be influenced by some basic practical considerations at least as much as your expertise in statistics, study design, or SQL. Here are hard-learned lessons from working in the field. The top areas to emphasize are:

1. Data source validation - So, you have a new data source, table, analysis, etc, learn how to check for common errors and data issues through simple validation techniques
   1. Calculate a well known number and match it against a trusted source
   2. Use temporary tables for each step of an analysis and check row counts as you go to prevent errors
   3. Build a list of checkable things that should be true about your new data source, then write queries to check them (e.g. no values below zero, values lie in a reasonable range, number of values makes sense given the number of users represented, etc)
2. Data interpretation skepticism - the first questions to ask about a new data source or a metric movement, are what do these data actually contain, how were they made and what do they include. Often the important fact about the data is some issue with the logging, or some secular change (new internet pricing, new competitor, new device type, holiday), or some artificial filter (only active users were surveyed) and not a meaningful movement in the metric we want to study. Full clarity about exactly what the data is, gets us many of the most important answers. Here is an example of the questions you might ask yourself when you see significant changes in your metric:

**Technical issues:**

* Is the metric movement (decrease/increase) just a logging issue? In other words was there any recent code release that could have had logging bugs (duplicated logs or logs missing completely/ in certain scenarios)?
* Is the hour of the first significant movement in metric correlates with any recent code release to production? If so, can you try to identify if all users are affected or certain segments only (that will help developers to find the problem)?
* Is there any specific OS / browser influenced? Was there any new release of the new OS or Browser version?
* Was there any change in the data collected to the tables used in SQL? Unexpected null values? Pipeline changes release with potential bugs? New types of dimensions values not existed before that gets filtered?

**Non-technical issues:**

* Was there any recent launch of competing features?
* Can it be a possible fraud attack?
* Is the hour of the first significant movement in metric correlates with any non-technical recent change performed by the team: user acquisition campaign? Change in settings of any other tools involved?
* Seasonality or special event (can also be for a specific country only)?
* Is it one piece of content that generated most of the increase in KPI?
* Was there any change in users distribution that could influence the overall metric?

1. People’s cognitive biases and agendas - sometimes you can have the perfect analysis that shows a clear actionable path forward, but your organization won’t adopt it. Why not? Because they aren’t being rational. This is the norm, not an exception. People may be biased toward their vision, their assumptions, confused about how to interpret the data. Explaining the consequences of following a certain course of action as clearly as possible, can help convince people. It’s also important to find allies in the business leadership who will champion adoption of the data analyst’s findings.

**7. User behavior**

When engineers build apps, they are creating an experience for real life people to try. The users’ responses will not be random, they will follow rules of human psychology.   
Some knowledge of how people are wired, how they think, and how they respond to stimuli is very important in order to propose product changes that have good odds of success. It also helps in figuring out the mechanics behind why a certain change drove metrics up or down - why did users respond as they did?

[Thinking Fast and Slow - Prof Daniel Kahaneman](https://www.youtube.com/watch?v=CjVQJdIrDJ0) - 1 hour video summary of his book

*Notice: the copyright for the works referenced belong to their respective owners.*