Analyzing the NYC Subway Dataset

Student Notes | Project Review

Does Not Meet Specifications

Communication

SPECIFICATION

Analysis done using methods learned in the course is explained in a way that would be understandable to a student who has completed the class.

MEETS SPECIFICATION

Reviewer Comments

Vernacular is consistent with the contents of the class.

SPECIFICATION

The answers are a well-formed summary of the analyses and do not leave out important information (i.e. fully answering the question).

DOES NOT MEET SPECIFICATION

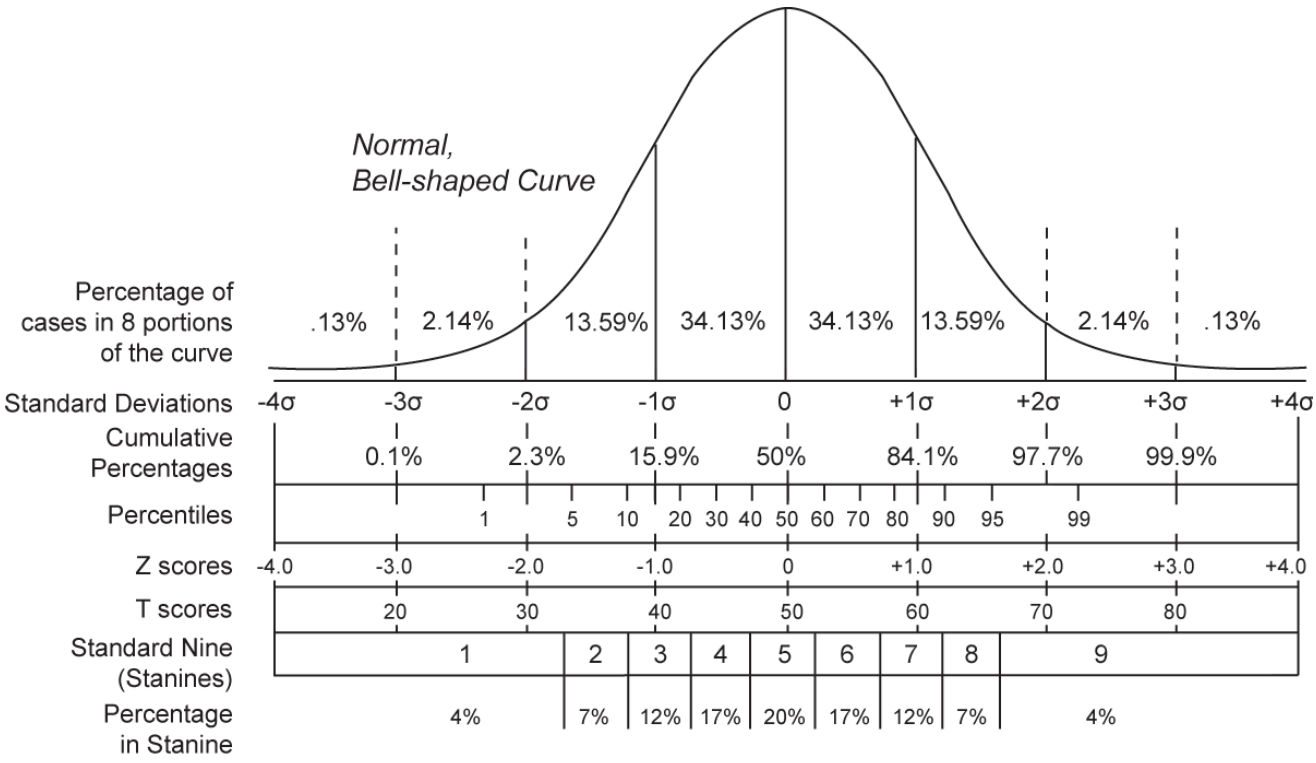
Reviewer Comments

Justifying Mann-Whitney

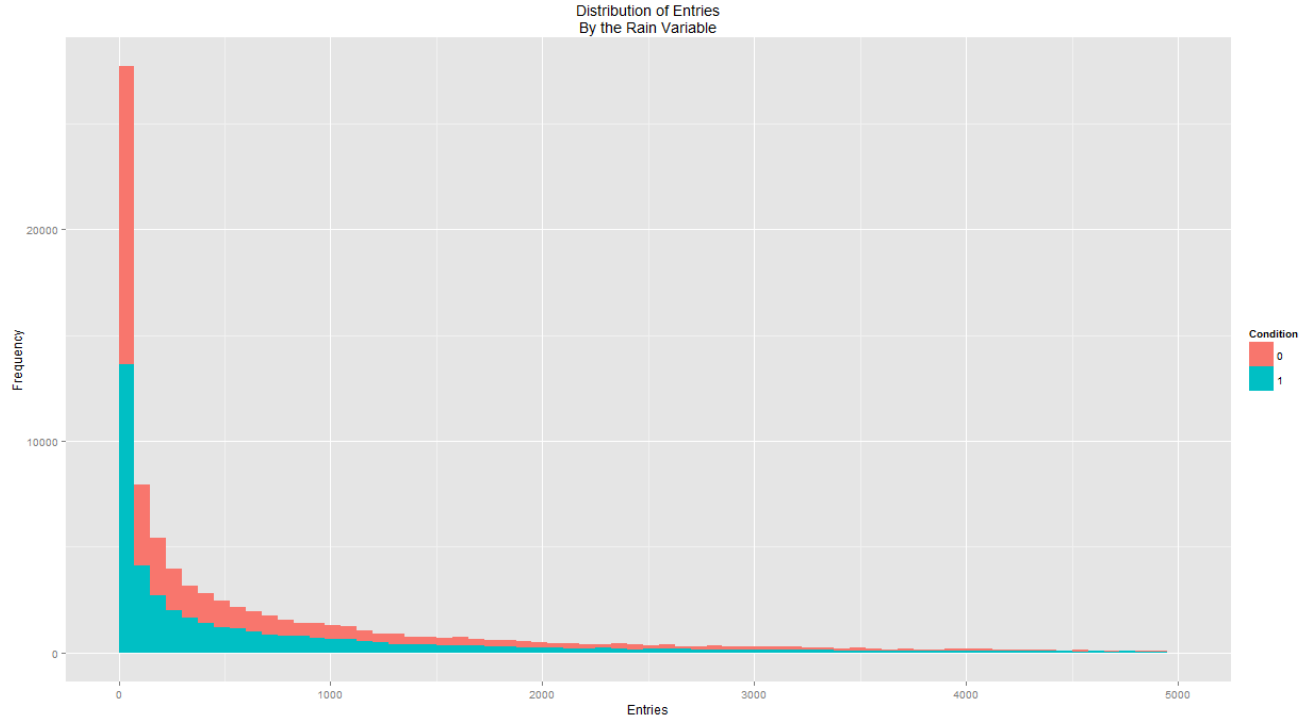
ANS : Because we do not know the data is droawn from any particular underlying probability distribution, two samples come from the same population against an alternative hypothesis, especially that a particular population tends to have larger values than the other .

We require a better justification for using the Mann-Whitney over the T-test. With the current justification, there would be no reason to even use the T-test and assessments of distribution would not be needed. For our dataset, we may not know what the distribution is, but we do know what it isn't. Please see below for clarification.

Below is the **Normal Distribution**



This is the histogram in **Section 3.1**. We required this part of the project in order to show the distributions of the `ENTRIESn_hourly` variable.



- Part of the assumption of parametric tests like the T-test is that the distributions are normal.
- As you can see, the distribution is **not normal**.
  This is the justification we should be using.
- Another measurable way to determine *Normality* is to use a normality test like the Shapiro-Wilk Test.
  - The Shapiro-Wilk is a statistical test that is performed the same way we have done with T-tests and Mann-Whitney Tests except the null hypothesis it that the sample comes from a normally distributed population. The alternative is that it does not.
  - https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

Please expand on the meaning of the R-squared

It is great that you have interpreted the value of R-squared, however, we must discern if you understand what the concept of R-squared is.

- Define what the R^2 means (in general terms so we know you understand the term) https://en.wikipedia.org/wiki/Coefficient_of_determination
- Apply that definition to the dataset; meaning - reword the definition so it applies to the data we have
- (OPTIONAL) Examining the residuals is a great way to justify the appropriateness of our model. I discuss this in-depth in a different area of the evaluation. It is optional, but **highly** recommended for students to familiarize themselves with.

What is your Conclusion?

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

From the null hypothesis testing rain has significant effect on Houly_entries.

From rain's coeffiecent in linear regression we can concluse that "The people do less ride when it is raining."

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

From the null hypothesis testing results with 2 samples rain and with out rain return p value = it mean that the possibility of different has occured by chance is about 0.0002741% so we assume that the rain has significantly effect on Houly_entries.

From the rain coeffiecent of linear regression is about -71.02 so we assume that the rain has effect on Houly_entries in negative way.

You have stated the results of your analysis, but have not answered the question - "do more people ride the NYC subway when it is raining or when it is not raining?" You have contradictory evidence. How would you justify and answer either way? Or if it is inconclusive, why?

1. Statistical Test
    - Compare the relative means of the two conditions
    - What is the result of the hypothesis test http://blog.minitab.com/blog/understanding-statistics/things-statisticians-say-failure-to-reject-the-null-hypothesis
2. Regression Analysis
    - What do the coefficients tell us about the features? http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients
    - If Dummy-variables are used, how are their coefficients compared to the weather variables' coefficients?
    - How valid is our Regression analysis?
        - To determine this, one might to an analysis of the Residuals..
        - Then describe the appropriateness of the model **with evidence**. One way to that is to examine the distributions of the residuals by viewing the residuals in a histogram. Be wary of long tails with high values when viewing the residual histogram - these typically mean we have some large errors and gives us ground to question the model. Here are a couple of methods to help....
          Normality of Residuals with Probability Plot
          http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html
          QQ plot

http://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

Quality of Visualizations

SPECIFICATION

Plots depict relationships between two or more variables.

MEETS SPECIFICATION

SPECIFICATION

All plots and data are of the appropriate type.

MEETS SPECIFICATION

SPECIFICATION

All plots are appropriately labeled and titled. Plot is given an appropriate title. X-axis and y-axis are appropriately labeled. Visual cues (colors, size, etc) are easy to distinguish. It is clear what data are represented.

DOES NOT MEET SPECIFICATION

Reviewer Comments

Please include a short description interpreting each plot.

Quality of Analysis

SPECIFICATION

When using statistical tests and linear regression models, the choice of test type and features are always well justified based on the characteristics of the data.

MEETS SPECIFICATION

Reviewer Comments

Great work assessing the *hour* variable as a categorical variable. Time-series is a tricky variable. In situations where one is concerned about how the passage of time affects the outcome variable, time-series variables can be treated as regular non-dummy variables. In this instance, however, we are not concerned about the passage of time as much as we are concerned about **what** time it is and how *that* affects the outcome variable. In this case, the time-series variable is, in fact, a categorical variable. This means the *hour* variable should technically be used as a dummy-variable with each instance in time having its own column and values indicating if it is that time or not.

SPECIFICATION

Statistical tests and linear regression models are described thoroughly, and the reasons for choosing them are articulated clearly.

MEETS SPECIFICATION

SPECIFICATION

The use and interpretation of statistical techniques are correct.

DOES NOT MEET SPECIFICATION

Reviewer Comments

Incorrect Hypotheses for Mann-Whitney

Currently, the symbol μ is used in the description of the null for the Mann-Whitney. Since μ is used to denote population mean, this would be an incorrect application. The Mann-Whitney is not concerned with means - this is why it is robust against outliers and unknown/non-normal distributions. Please look at the link below and particularly #3 and #4 for an appropriate statement of the hypotheses.
https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#Assumptions_and_formal_statement_of_hypotheses

---

One-tailed P-value reported

```
We reject the null hypothesis (μ(with rain) ≠ μ(without rain))
p-values is 0.000002741069571243749558478279559903256767938728444278240203857421875000000000000000000000
with rain mean is 2028.1960354721
without rain mean is 1845.5394386644
```

The default returned p-value for the Mann-Whitney is one-tailed. Multiply the highlighted value by 2 in order to get the two-tailed (since you are seeking the two-tailed value in 1.1).

SPECIFICATION

All conclusions are correctly justified with data.

MEETS SPECIFICATION

Reviewer Comments

RESIDUAL (MODEL ERROR) ANALYSIS IS A VITAL SKILL TO HAVE IN JUDGING THE PERFORMANCE AND APPROPRIATENESS OF A MODEL. THE FOLLOWING IS OPTIONAL BUT I HIGHLY RECOMMEND FAMILIARIZING YOURSELF WITH THIS MATERIAL AS IT IS NOT ONLY APPLICABLE TO MANY OF THE QUESTIONS IN THE PROJECT, BUT IS CRUCIAL TO YOUR FUTURE AS A DATA SCIENTIST.

Interpreting the Residuals (Optional)

Examining the residuals (errors) of a model can give us some great insights to the behavior of the dataset. In the following I have given a simple rundown on how to examine them. As you will see, it can be used to justify many of the sections in the project.

THE FOLLOWING MATERIAL IS A GREAT WAY TO JUSTIFY THE CONCLUSIONS MADE IN **SECTION 2.6**, **SECTION 4**, AND **SECTION 5** AND IS HIGHLY RECOMMENDED TO FAMILIARIZE YOURSELF WITH IT.

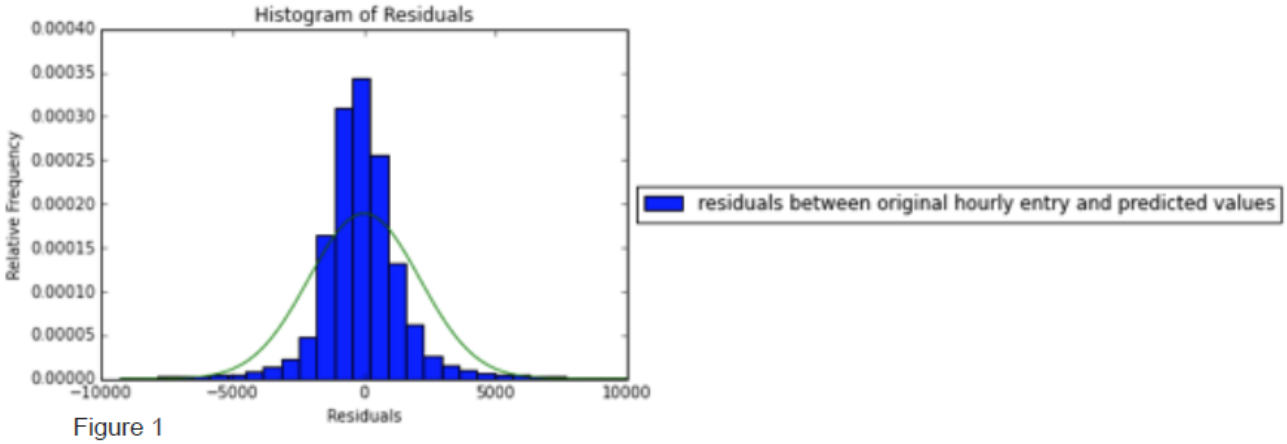


Figure 1

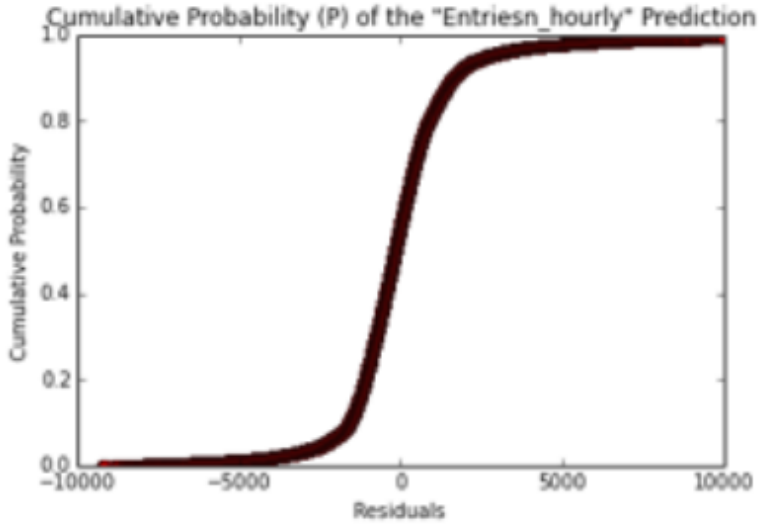
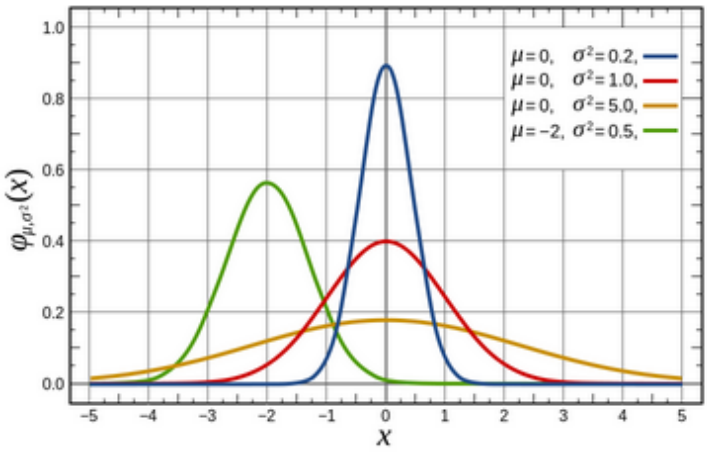


Figure 2

- Above, in *Figure 1*, I have included a version of our residuals with an overlay of the normal distribution, also known as the **P**robablity **D**ensity **F**unction **(PDF)**. Below that, in *Figure 2*, is the **C**umulative **D**ensity **F**unction **(CDF)** of our residuals. Both are great ways to visually determine distributions.
- Below are examples of both these functions with different **means** and **standard deviations**.

This is the **Probability Density Function**.

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



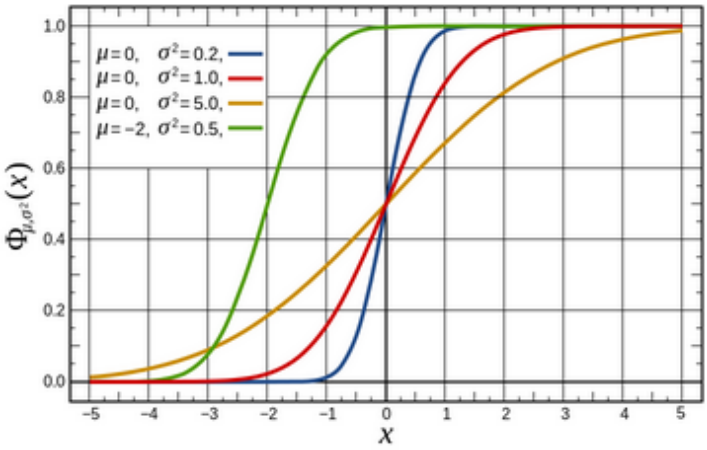
"Normal Distribution PDF" by Inductiveload - self-made, Mathematica, Inkscape. Licensed under Public Domain via Commons

If $\mu = 0$ and $\sigma = 1$, the distribution is called the **standard normal distribution** or the **unit normal distribution** denoted by $N(0,1)$ and a random variable with that distribution is a **standard normal deviate**.

- The curve with the **red line** represents the normal distribution
- One can visually tell obviously non-normal distributions, but is there are way to test it measurably?
    - **Shapiro-Wilk**
      https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
      The Shapiro-Wilk Test is a common statistical test for checking normality they same way we test our hypothesis for T-tests and Mann-Whitney U-tests. The **Null** is that the population is normally distributed and we test to reject or fail to reject that null.

This is the **Cumulative Distribution Function**

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$



"Normal Distribution CDF" by Inductiveload - self-made, Mathematica, Inkscape. Licensed under Public Domain via Commons

- The **red line**, again represents the normal distribution.

The Key idea I want to convey is that if we were to examine our residuals and compare their distributions to the Normal Distributions above, we would see that the residuals are not normally distributed. That, coupled with the fact that the residual histogram has very long-tails with extremely high absolute errors, means that

our model is actually NOT a good fit for the dataset.

SPECIFICATION

No incorrect conclusions are drawn from the data.

MEETS SPECIFICATION

SPECIFICATION

Some shortcomings of the dataset and statistical tests or regression techniques used are appropriately acknowledged.

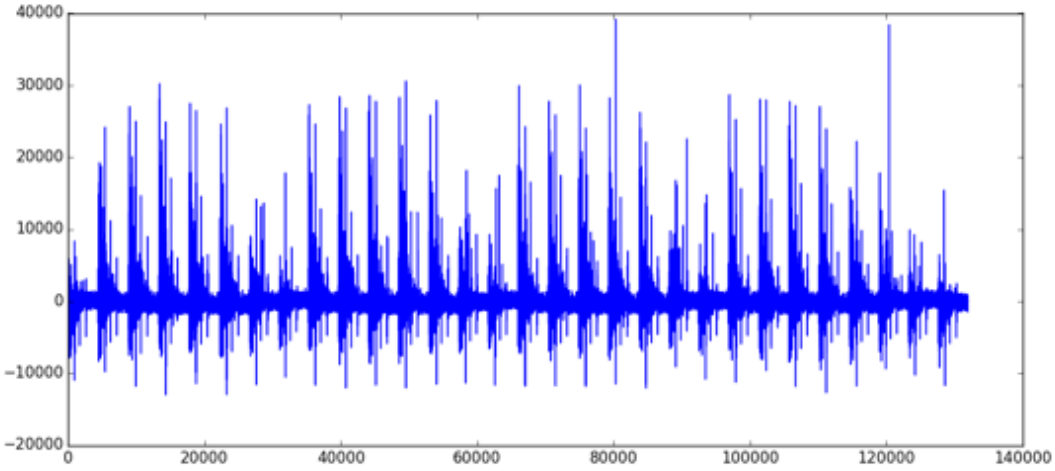
DOES NOT MEET SPECIFICATION

Reviewer Comments

More discussion required for the potential Shortcomings

Please expand on this area. Here are some general ideas. Some you may have already mentioned but I like to keep to maintain continuity.

1. **Dataset**
   - Consider the scope of the data. It spans a month. Do you believe that is long enough?
   - Because there are many variables included in the dataset that might be very closely related, such as minimum, mean and maximum temperature, it may be difficult to disentangle the effects of such similar features and we may run the risk of problems with multicollinearity, which can cause some linear regression algorithms to give incorrect results.
     http://en.wikipedia.org/wiki/Multicollinearity
2. **Statistical Test**
   - You can talk about the fact that the Statistical Test is only comparing the differences between the conditions of one feature; rain, when clearly other are other variables that seem to affect the ridership even more.
3. **Regression Model**
   Do you think that a linear model is appropriate in this context? Why? You could simply extend to 5.1 the reasoning regarding the residuals I proposed earlier. In addition it might be interesting to plot the residual per data point, some interesting patterns might emerge to help understand why a linear model might not be the best choice for this problem.
   As I have given hints in the prior section, examining the residuals of a regression model is one of the best ways to determine the effectiveness of the model. You can use any of the techniques above. By merely plotting the difference between predictions and actual values (residuals) you will, most likely, see that the residuals follow a cyclical pattern. If so, that might prove that some non-linearity in the data should be addressed by designing a non linear model. The code is really simple and looks like this: import matplotlib.pyplot as plt
   plt.plot(data - predictions) plt.show().
   The following is an example of what you might see (randomly created).



You may see a plot similar to this that maybe shows more of a cyclical/higher degree shape of the residuals. This could indicate that the outcome does not respond linearly to the features and a linear model might not be the best fit.

Download project

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

How satisfied are you with this feedback?

Resubmit Project

include the link to this review.

NANODEGREE PROGRAMS

Front-End Web Developer
Full Stack Web Developer
Data Analyst
iOS Developer
Android Developer
Intro to Programming
Tech Entrepreneur

STUDENT RESOURCES

Blog
Help & FAQ
Catalog
Veteran Programs

PARTNERS & EMPLOYERS

Georgia Tech Program
Udacity for Business
Hire Nanodegree Graduates