




Journey To Be A Data Scientist



A famous saying goes like this....

A Data Scientist should know statistics more than a computer programmer and Computer Programming more than a Statistician .

- *THIS COURSE MODULE IS DESIGNED FOR ABOVE SAYING.*
- *YOU LEARN , WHAT YOU CHOOSE.*
- *YOU STOP AND EXIT ON YOUR CHOICE .*
- *JOIN US BACK , ON YOUR WISH.*
- *PAY ONLY FOR JUST WHAT YOU WANT TO LEARN.*

I am not a TRAINER.

► Yes , you hear me right !

I am your Learning Companion , your Study Buddy and Your Coach , who would be with you throughout and beyond this journey.

Unlike Trainers, Who Is Concerned With How Much Time You Put In. A Learning Companion is Concerned With How Much You Put Into The Time.

Data Science Journey is long , everlasting but Fun. I need your motivation.

Your Skill Showcase is Your Certificate.

There is no debate , that just getting a Certificate has become so ordinary these days. What employers are looking , is not your certificate, rather your skills. No body gets impressed by certificates anymore, but they do get if they see your actual work.

Read this article about this myth.

<https://www.dataquest.io/blog/data-science-certificate/>

Having said that.....

- Throughout our course, We will focus on building Projects, Assignments, White Papers in a presentable form, so that you can upload them on your github , Kaggle profiles.
- Share these links on your resume , Linkindln and public profiles to actually impress your potential employers by your work and not by certificates.

**Get
Started**



Module I : Getting Started (10 Hours)

R Basic Concepts

- ▶ Statistical Computing and R
- ▶ R Basics
- ▶ Functions and Distributions
- ▶ Creating Sequences

Control Flow and Functions

- ▶ Expressions and Compound Expressions
- ▶ Loops
- ▶ Functions
- ▶ More Examples
- ▶ Vectorised Computation

Matrices

- ▶ Creating Matrices
- ▶ Matrix Subsets
- ▶ Simple Computations with Matrices
- ▶ Matrix Operations

Statistical Data

- ▶ Categorical Data
- ▶ Data Frames
- ▶ Reading and Printing Data
- ▶ Low-Level Data Input
- ▶ Connections

Graphics

- ▶ Introduction
- ▶ Low-Level Graphics and Drawing Functions
- ▶ Choosing Colours to Fill Areas
- ▶ Eye-Catching Visual "Tricks"
- ▶ Layouts and Margins
- ▶ Using par
- ▶ Layouts and Plot

Module I : Getting Started

Scoping

- ▶ Variable Scope
- ▶ Scoping Rules
- ▶ Closures
- ▶ Why Are Closures Useful

Its time for some Assignments.....

- ▶ Assignment 1
- ▶ Assignment 2

Lets Accelerate now

- ▶ Okay ! Now we have learnt Basics of R Programming Language. However this is not enough for Enterprise level Data Analysis Problems Solving.

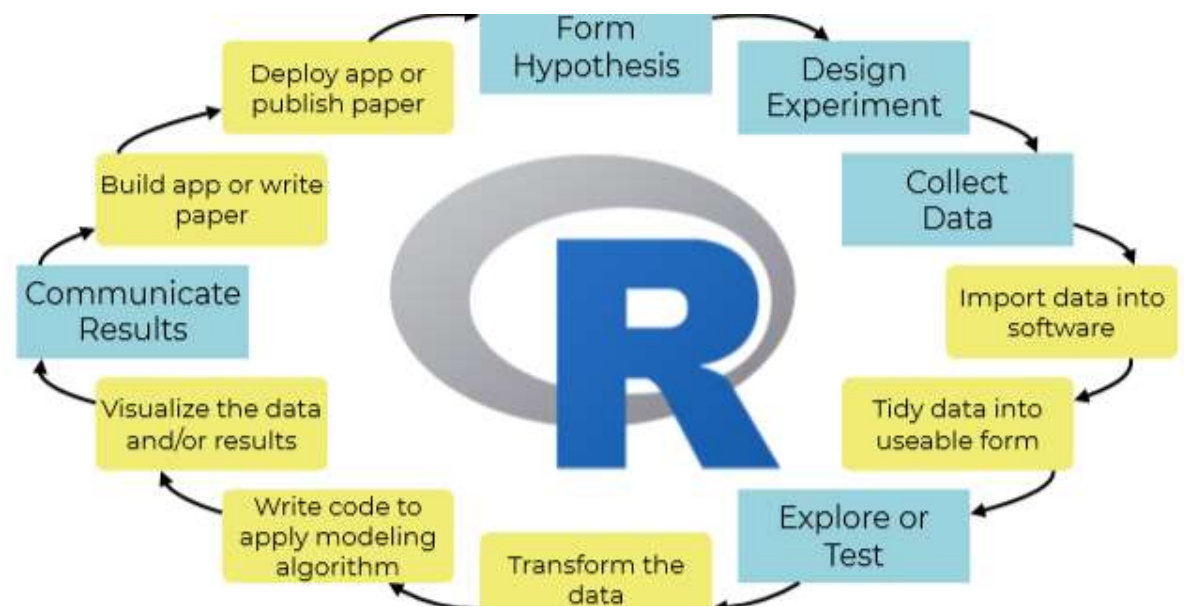
We must make our life easier....So there are bunch of R Packages , already designed for Quick Data Manipulation , Data Exploration etc.

Tidyverse package makes data science faster, easier and more fun with R.

- ▶ Tidyverse is a coherent system of packages for data manipulation, exploration and visualization that share a common design philosophy.
- ▶ Tidyverse packages are intended to make statisticians and data scientists more productive by guiding them through workflows that facilitate communication, and result in reproducible work products.

Module II : Getting Ready (10 Hours)

- ▶ Data Visualization with **ggplot2**
- ▶ Data Transformation with **Dplyr**
- ▶ Exploratory Data Analysis
- ▶ Tidy Data with **tidyr**
- ▶ **Pipes** and **Functions**
- ▶ **Communicate** with R Markdown



Where is the Science ?

Excuse Me !!!

We have just Learnt , how to use tools , but where do I have to hit it ?

Now lets learn the Science !

Hmm...Lets Start with Statistics and Move to Machine Learning.

Module III : Basic Statistics (15 Hours)

INTRODUCTION

- ▶ What Is Statistics?
- ▶ Statistics in Our Everyday Life
- ▶ Statistics in Aid of Scientific Inquiry
- ▶ Two Basic Concepts—Population and Sample
- ▶ The Purposeful Collection of Data
- ▶ Statistics in Context
- ▶ Objectives of Statistics
- ▶ Using Statistics Wisely

ORGANIZATION AND DESCRIPTION OF DATA

- ▶ Main Types of Data
- ▶ Describing Data by Tables and Graphs
- ▶ Categorical Data
- ▶ Discrete Data
- ▶ Data on a Continuous Variable
- ▶ Measures of Center
- ▶ Measures of Variation

PROBABILITY DISTRIBUTIONS

- ▶ Introduction to Probability
- ▶ Probability of an Event
- ▶ Methods of Assigning Probability
- ▶ Equally Likely Elementary Outcomes— The Uniform Probability Model
- ▶ Probability As the Long-Run Relative Frequency
- ▶ Event Relations and Two Laws of Probability
- ▶ Conditional Probability and Independence
- ▶ Random Variables
- ▶ Probability Distribution of a Discrete Random Variable
- ▶ Expectation (Mean) and Standard Deviation of a Probability Distribution
- ▶ Successes and Failures—Bernoulli Trials
- ▶ The Binomial Distribution
- ▶ The Normal Distribution—Its General Feature
- ▶ Central Limit Theorem

Module III : Inferential Statistics (25 Hours)

DRAWING INFERENCES FROM LARGE SAMPLES

- ▶ Introduction
- ▶ Point Estimation of a Population Mean
- ▶ Confidence Interval for a Population Mean
- ▶ Testing Hypotheses about a Population Mean
- ▶ Inferences about a Population Proportion

SMALL-SAMPLE INFERENCES FOR NORMAL POPULATION

- ▶ Student's t Distribution
- ▶ Inferences about mean — Small Sample Size
- ▶ Confidence Interval for mean
- ▶ Hypotheses Tests for mean
- ▶ Relationship between Tests and Confidence Intervals (The Chi-Square Distribution)

COMPARING TWO TREATMENTS

- ▶ Independent Random Samples from Two Populations
- ▶ Comparing Two Population Proportion
- ▶ Randomization and Its Role in Inference

SIMPLE LINEAR REGRESSION

- ▶ Introduction
- ▶ Regression with a Single Predictor
- ▶ Straight-Line Regression Model
- ▶ The Method of Least Squares
- ▶ The Sampling Variability of the Least Squares Estimators— Tools for Inference
- ▶ Important Inference Problems
- ▶ Inference Concerning the Slope
- ▶ Inference about the Intercept
- ▶ Estimation of the Mean Response for a Specified x Value
- ▶ Prediction of a Single Response for a Specified x Value
- ▶ The Strength of a Linear Relation.

Module III : Inferential Statistics

▶ **MULTIPLE LINEAR REGRESSION**

- ▶ Estimating the Regression Coefficients.
- ▶ Model Assumptions.
- ▶ Diagnostics.
- ▶ Data Problems - High Leverage Points, Outliers and Multicollinearity.
- ▶ Methods and Algorithms to detect data Problems.
- ▶ Steps to Remove Data Problems.
- ▶ A Lack of Fit Test.
- ▶ Interactions and ANOVA

▶ **MODEL SELECTION**

- ▶ Akaike's information criterion (AIC)
- ▶ Bayesian information criterion (BIC).
- ▶ Model Selection Using Loss Function
- ▶ Ridge and LASSO Regression
- ▶ Cross Validation , K Fold and LOOCV

Its Time for some Assignments

- ▶ Assignment 3
- ▶ Assignment 4
- ▶ Assignment 5

Are you interested for More Assignments ? We do have bunch of those, created by Top data scientist and Universities.

We will also share Best Books and Materials to accelerate your learning.

Hey , Don't get Tiered !! Remember motivation is Key in this journey.

You know, By this time you are ready for 60% data analysis. But you are still not a Data Scientist.

Time for Project

Yes. After all you must showcase your skills to potential employer.

A project means learning by doing: To consolidate what you learnt so far. To pour your ideas to solve problems. To enhance your project by using what you gonna learn next.

All interview starts....What projects you did in Data Science field.....

I hope you got , what I mean now !!

Module IV : Gearing Up for Predictive Modeling

(5 hours)

- ▶ Models
- ▶ Types of Models
- ▶ The Process of Predictive Modeling

Performance metrics

- ▶ Assessing regression models
- ▶ Assessing classification models
- ▶ Assessing binary classification models

The process of predictive modeling

- ▶ Defining the model's objective
- ▶ Collecting the data
- ▶ Picking a model
- ▶ Preprocessing the data
- ▶ Exploratory data analysis
- ▶ Feature transformations
- ▶ Encoding categorical features
- ▶ Missing data
- ▶ Outliers
- ▶ Removing problematic features
- ▶ Feature engineering and dimensionality reduction
- ▶ Training and assessing the model
- ▶ Repeating with different models and final model selection
- ▶ Deploying the model

Module IV : Machine Learning with R (20 Hours)

Logistic Regression

- ▶ Classifying with linear regression
- ▶ Introduction to logistic regression
- ▶ Generalized linear models
- ▶ Interpreting coefficients in logistic regression
- ▶ Assumptions of logistic regression
- ▶ Maximum likelihood estimation
- ▶ Assessing logistic regression models
- ▶ Model deviance
- ▶ Test set performance
- ▶ Example : Predicting Heart Disease.

Neural Networks

- ▶ The biological neuron
- ▶ The artificial neuron
- ▶ Stochastic gradient descent
- ▶ Gradient descent and local minima
- ▶ The perceptron algorithm
- ▶ Multilayer perceptron networks
- ▶ Training multilayer perceptron networks
- ▶ Evaluating multilayer perceptron for regression
- ▶ Receiver operating characteristic curves
- ▶ Example : Predicting the energy efficiency of buildings

Module IV : Machine Learning with R

- ▶ **Support Vector Machine**
- ▶ **Tree Based Machine Learning Algorithms**
- ▶ **Random Forest**
- ▶ **Ensemble Methods**
- ▶ **Clustering and Segmentation**

Optimizing Your Model

- ▶ K-Fold Cross Validation
- ▶ Tuning Model Parameters

Time to Review Your Project

Module V : Fast Analytics with Tableau

Introduction

- ▶ Introduction to Tableau and an overview of the different versions
- ▶ Installing Tableau Desktop
- ▶ Tableau Help and online resources

Working With Tableau

- ▶ Understanding Tableau User Interface
- ▶ Exploring Tableau File Types
- ▶ Understanding Green and Blue Pills
- ▶ Working with Available Data Sources
- ▶ Working with extracts instead of live connections

Deep diving with Data and Connections

- ▶ Working with Excel Data Interpreter
- ▶ Learning how to Split fields
- ▶ Pivoting Data
- ▶ Understanding Metadata and sharing data source connections
- ▶ Filtering data from your data source

Module V : Fast Analytics with Tableau

Creating Charts

- ▶ The Show Me Feature
- ▶ Crosstabs and Heat Maps
- ▶ Using Bar, Stacked Bar and Side-by-side Bars
- ▶ Pie Charts
- ▶ Line and Area Charts
- ▶ Working with Packed Bubble
- ▶ Using Treemaps
- ▶ Creating a Basic Scatter Plot
- ▶ Creating a Basic Map

Adding calculations to your workbook

- ▶ Introduction to Calculations
- ▶ Understanding Basic calculations
- ▶ Understanding String Calculations
- ▶ Learning about Boolean, If-Then calculations, and Case statements
- ▶ Understanding the Basics of Date Calculations
- ▶ Understanding Aggregation and Disaggregation
- ▶ Using calculations to add insight to your visualizations

Mapping data in Tableau

- ▶ Introduction to Mapping in Tableau
- ▶ Create a standard map view
- ▶ Dealing with map errors
- ▶ Customizing a standard map view
- ▶ Using filters in maps

Dashboards and Stories

- ▶ Introduction to Dashboards
- ▶ Understanding Dashboard Actions
- ▶ Understanding Dashboard formatting basics
- ▶ Understanding Workbook level formatting
- ▶ Assembling Your Dashboards Into A Story

The End Result of this course - That you create proper enterprise level interactive dashboard and publish it on free cloud environment, So that you can share that link to potential employer, linkedIn and Resume.

Talk To Us For More....

- ▶ Python For Data Science.
- ▶ Data Analytics and Machine Learning using Python.
- ▶ Deep Learning
- ▶ Fast Analytics Using Tableau
- ▶ Fast Analytics Using Power BI
- ▶ SQL
- ▶ Microsoft Azure : Machine Learning Studio
- ▶ SPSS