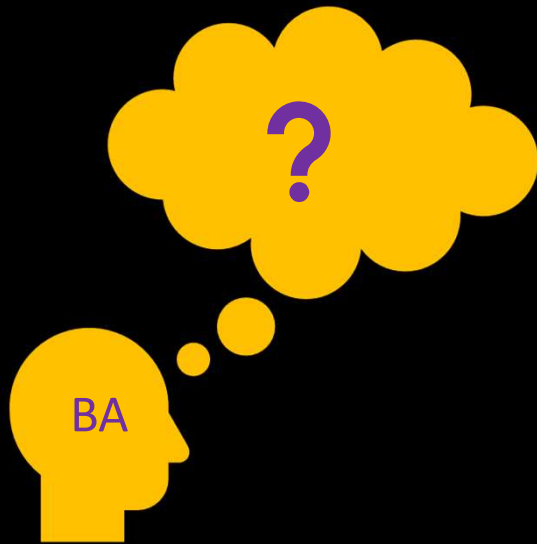


So You Want to
Become a Data
Scientist, Kinda?

What Comes Next ?

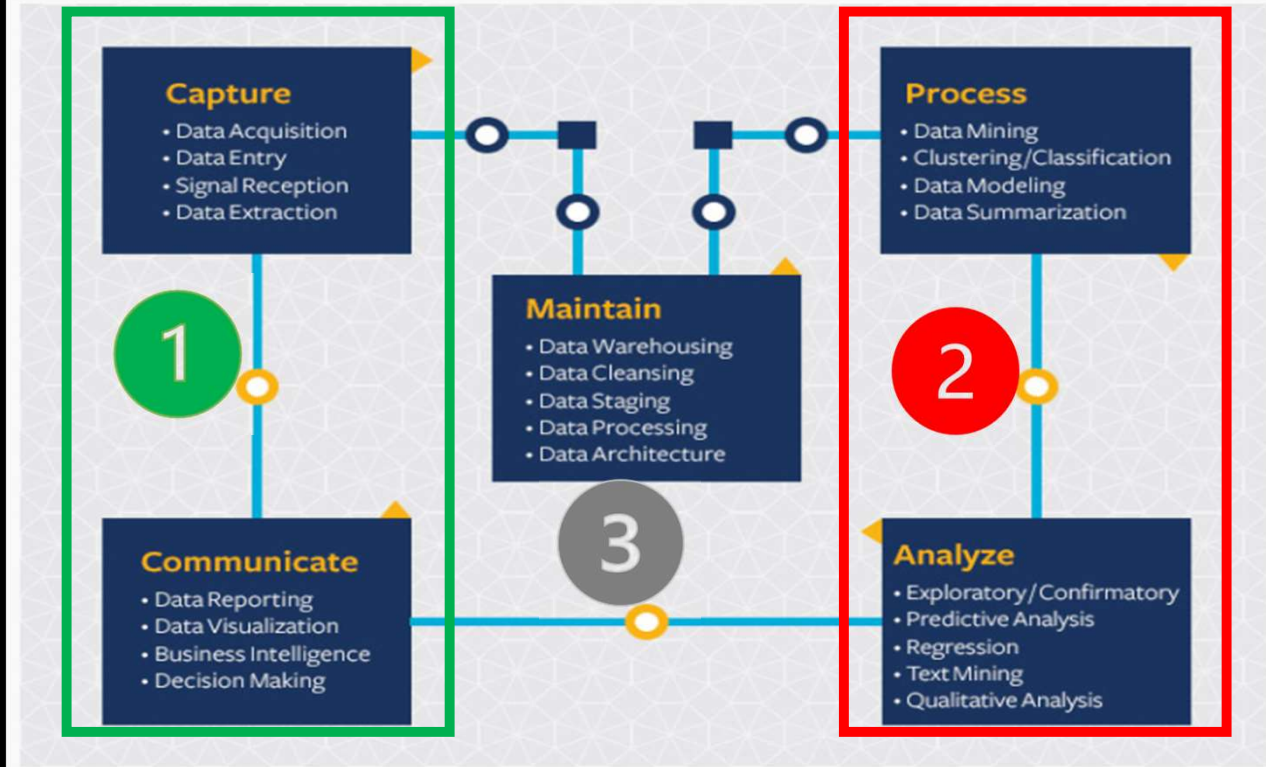
Beyond Analysis Conference
IIBA ATL Chapter
January 23, 2021





Traditional BA Tasks

The Data Science Life Cycle



What is Data Science ?

<https://ischoolonline.berkeley.edu/data-science/what-is-data-science/>

What is a Data Scientist ?

- “ Data scientists are a new breed of analytical data expert who have the technical skills to solve complex problems – and the curiosity to explore what problems need to be solved. ”



Ingredients:



Data



Be an “Insights Machine”



Domain Knowledge



Algorithms & Statistics



Processes



Communication

No Magic
Formula,
You
Define the
Roadmap

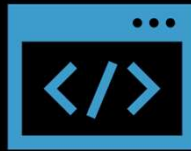


“If It’s to be, then it’s up to me.”

Agenda



Changing the BA's
Mindset



Show and Tell ... Key
Technologies



Transformation for BAs
and Practical Strategies



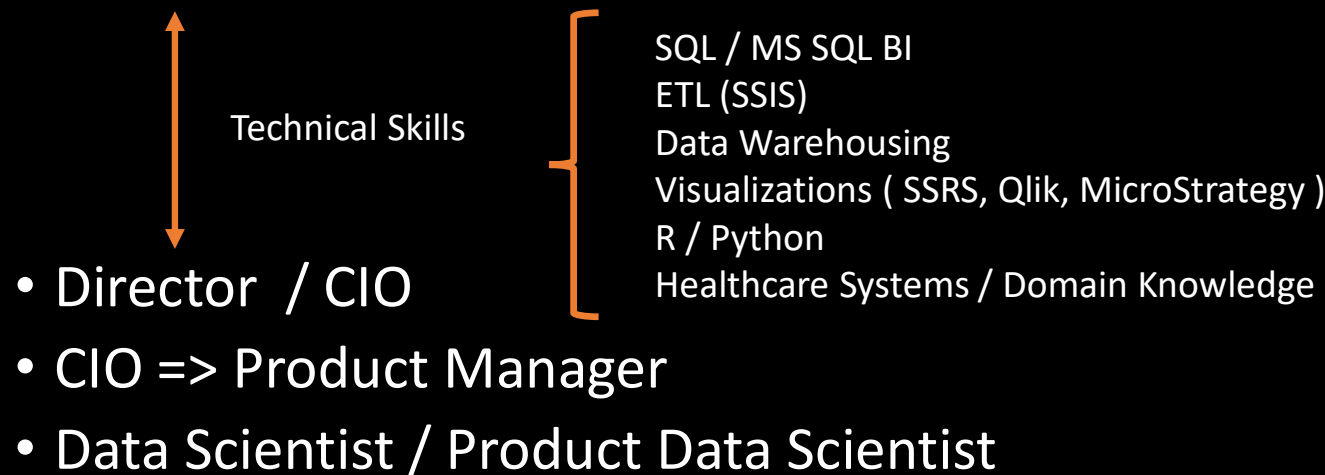
Changing the BA's Mindset

BAs 2010 vs. BAs 2021 Mindsets

Criteria	Old School: "Enterprise Support"	New School: "Data Centric"
Method	"Process Oriented"	"Data-Oriented"
Tools	1-2	Many & Variety of tools / data sets
Deliverable(s)	Traditional report	Innovators with data/game changer
Day-to-Day Focus	Across Enterprise	Functional / Team contributor
Autonomy	Low - Medium	High
Project Management	High & Collaboration	Self-Management

Career Transformation: Case Study

- Accounting Major => CPA => CFO
- CPA => Regulatory Expert (SEC / Sox / IT / Healthcare)
- Regulatory Expert => Entrepreneur / Consultant / Interim Exec
- Consultant => SQL BI Developer / Analyst



“Ingredients” to Transformation



INTELLECTUAL
CURIOSITY



PERSEVERANCE



PASSION



OPPORTUNITIES



PROBLEM SOLVER



ANALYTIC
ABILITIES

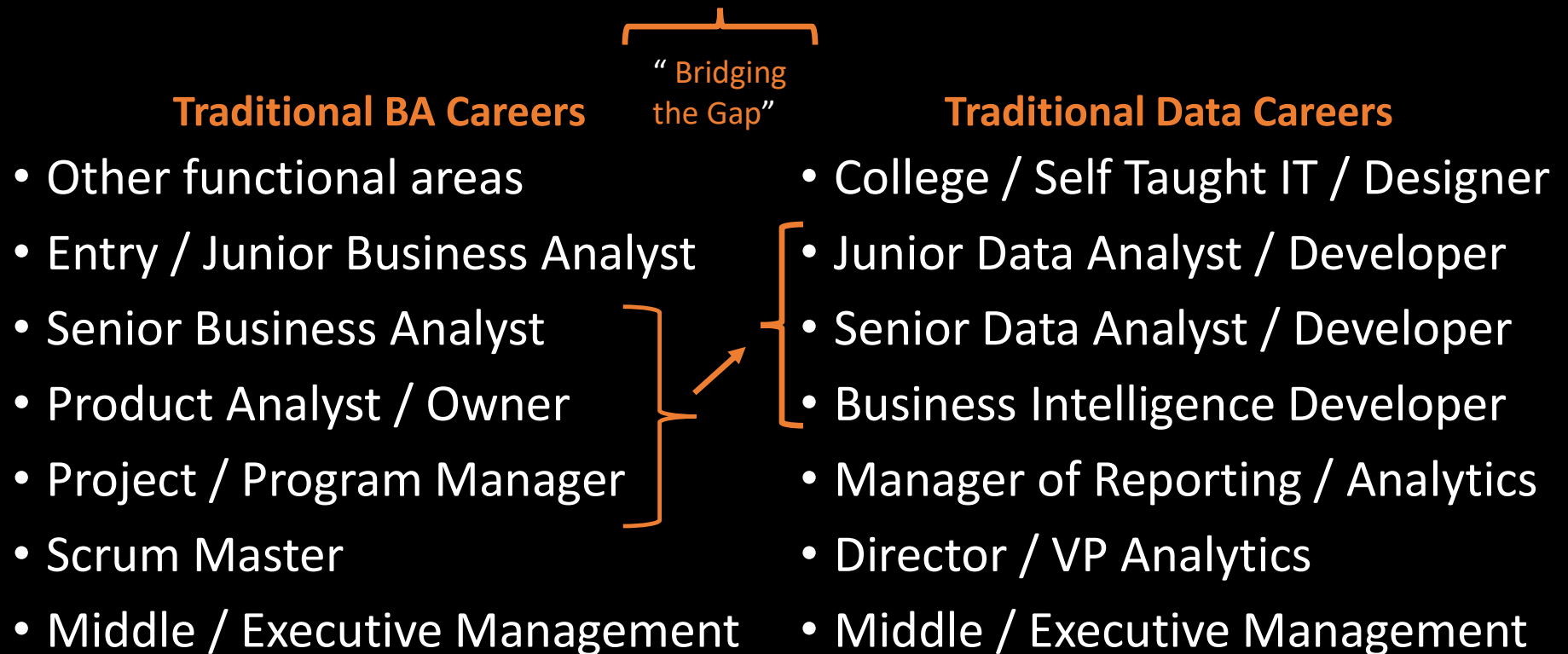


STRATEGIST &
TACTICIAN



“ UNDERDOG ”

Career Paths of BAs vs. Data Professionals



Leverage the Skills You Have



ANALYTICAL
MIND



DOMAIN
KNOWLEDGE



PROBLEM
SOLVING



CURIOSITY



LEADERSHIP /
INITIATIVE



INTERPERSONAL
SKILLS

Lessons Learned

Career change is hard, scary and frustrating.

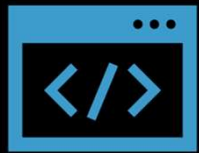
Understand your limits and strengths.

Technical skills are developed with practice, consistency and initiative.

Deadlines are great motivators, use them to your advantage.

There is no finish line, be lifelong learner.

Success is not binary, it's a continuum.

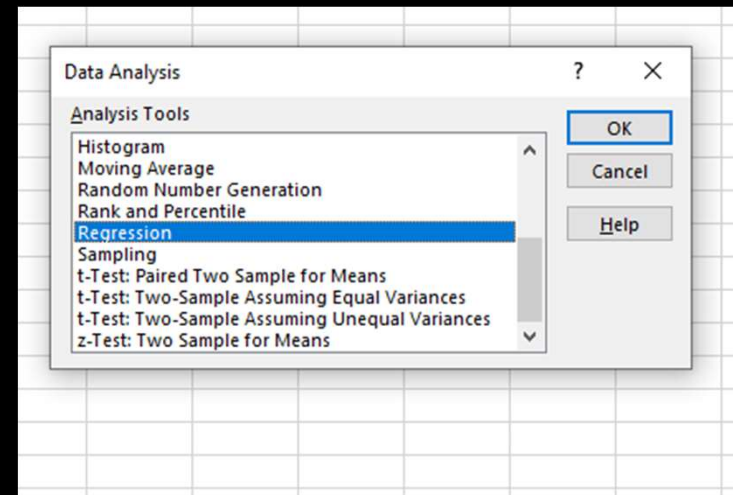
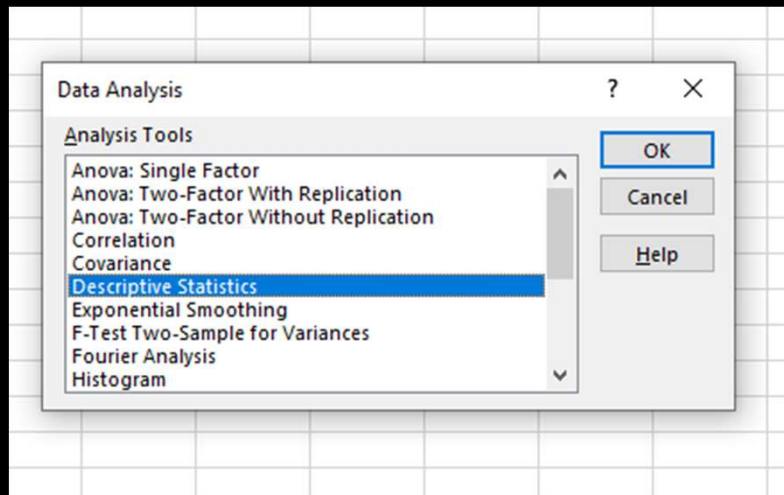


Show and Tell ... Key Technologies

Data Tools and Purpose

Tool	Familiarity	Sophistication	Purpose
Microsoft Excel	H	L - M	All Purpose
SQL Tools	L – M	M	Query data sets
E.T.L. / Data Warehouse	L – M	M – H	Extraction / Manipulation
Statistical / Development	L	M – H	“True” Analytics
Visualizations	M - H	L - M - H	Communication

Did You Know ? Excel used for Data Science !

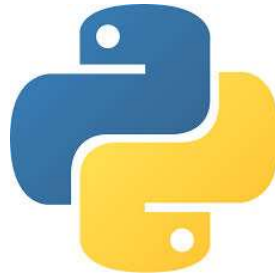


Using the **Data Analysis** Tab, you can perform basic statistical functions.





Data Warehouse Tools

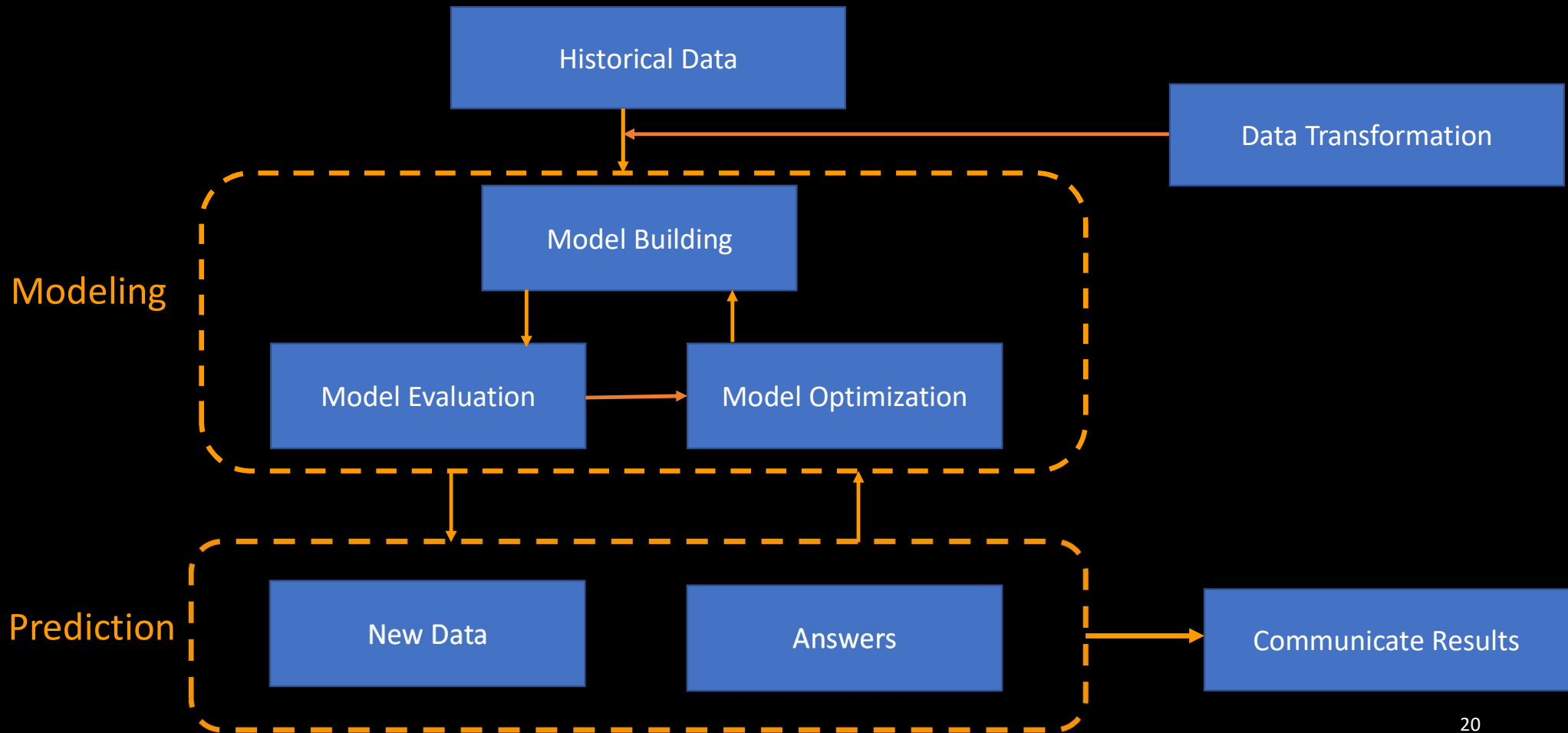


Statistical Development Tools



Integrated Development Environment Tools

Machine Learning Workflow – What D.S. does





Show and Tell



Example of Running RStudio IDE

The screenshot displays the RStudio IDE interface with the following components:

- Source Editor:** Contains R code for data analysis. The code includes comments about the source (Brett Lantz and Jay Roy), imports the `psych` library, sets device options, reads the `insurance.csv` file, and performs descriptive statistics like `summary`, `hist`, and `cor`.
- Environment:** Shows the loaded data frame `insurance` with 1338 observations and 7 variables: `age`, `sex`, `bmi`, `children`, `smoker`, `region`, and `expenses`.
- Console:** Displays the output of the executed R code, showing the summary of `insurance$expenses` and the histogram.
- Plots:** A histogram titled "Histogram of insurance\$expenses" is shown, with the x-axis labeled `insurance$expenses` and the y-axis labeled `Frequency`.

```
5 ## Original Example comes from book Machine Learning with R. Author: Brett Lantz and "tweaked" by Jay Roy.
6
7
8 ## Example: Using Linear Regression to Predicting Medical Expenses ----
9
10 ## Step 1: Import necessary libraries prior to data profiling exploring and preparing the data ----
11
12 ## You can obtain the libraries from https://cran.r-project.org/
13 library(psych)
14
15 dev.off()
16 c(1, 1, 1, 1)
17 par(mar = c(.5, .5, .5, .5))
18
19 ## library(stats) already included in base R and used to build your models.
20
21 ## Step 2: Data Profiling: Exploring data to obtain a better understanding of it.
22 ## This part of the Analytics process is called "descriptive statistics" ----
23
24 ## Read the csv file from kaggle.com into R.
25
26 insurance <- read.csv("insurance.csv", stringsAsFactors = TRUE)
27
28 ## Determine where your data sits?
29 getwd()
30
31 str(insurance)
32 dim(insurance)
33
34 # summary the medical expenses (charges) variable
35 summary(insurance$expenses)
36
37 # histogram of insurance charges - right tailed (skew)
38 hist(insurance$expenses)
39
40 # table of region
41 table(insurance$region)
42
43 # exploring relationships among features: correlation matrix
44 cor(insurance[c("age", "bmi", "children", "expenses")])
45
46
```

Console Output:

```
> insurance <- read.csv("insurance.csv", stringsAsFactors = TRUE)
> summary(insurance$expenses)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1122   4740   9382  13270  16640  63770
> hist(insurance$expenses)
```

Histogram of insurance\$expenses

insurance\$expenses	Frequency
0 - 10000	300
10000 - 20000	250
20000 - 30000	100
30000 - 40000	50
40000 - 50000	20
50000 - 60000	10

Case Study: Predicting Medical Costs

Scenario:

- You are the new Chief Data Scientist at “Nashville Health Plan”
- CEO requests you revitalize the profitability of this organization by identifying high-cost patients and predicting future costs.
- Your new team will present its findings at executive strategy meeting.

Questions:

1. How do you attack the problem?
2. What statistical tools could you use?
3. How do I explain this in a manner that my audience can consume?

Show and Tell: Python

Key Concepts:

- IDE – PyCharm vs Spyder vs. Jupyter Notebook
- How an IDE works and navigation in Pycharm
- Data Structures
- Modules
- Methods/Functions
- Objects
- Dot notation and attributes
- Python vs R programming (similarities and differences)

Example of Running Jupyter Notebook IDE

The screenshot displays the Jupyter Notebook IDE interface in a web browser. The browser's address bar shows the URL: `localhost:8888/notebooks/Documents/Data%20Mining%20for%20Bus%20Analytics/Python/Book%20Code/Untitled.ipynb?kernel_name=python3`. The Jupyter interface includes a top bar with the notebook title "Untitled - Jupyter Notebook", a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help), and a toolbar with icons for file operations and execution. The main area contains a code cell with the following Python code:

```
In [8]: %matplotlib inline
        from pathlib import Path
        import numpy as np
        import pandas as pd
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import r2_score
        from sklearn.linear_model import LinearRegression

        import matplotlib.pyplot as plt
        import os as os
```

Below the code cell, the output of the previous cell is shown:

```
In [9]: os.getcwd()
Out[9]: 'C:\\Users\\JayRoy\\Documents\\Data Mining for Bus Analytics\\Python\\Book Code'
```

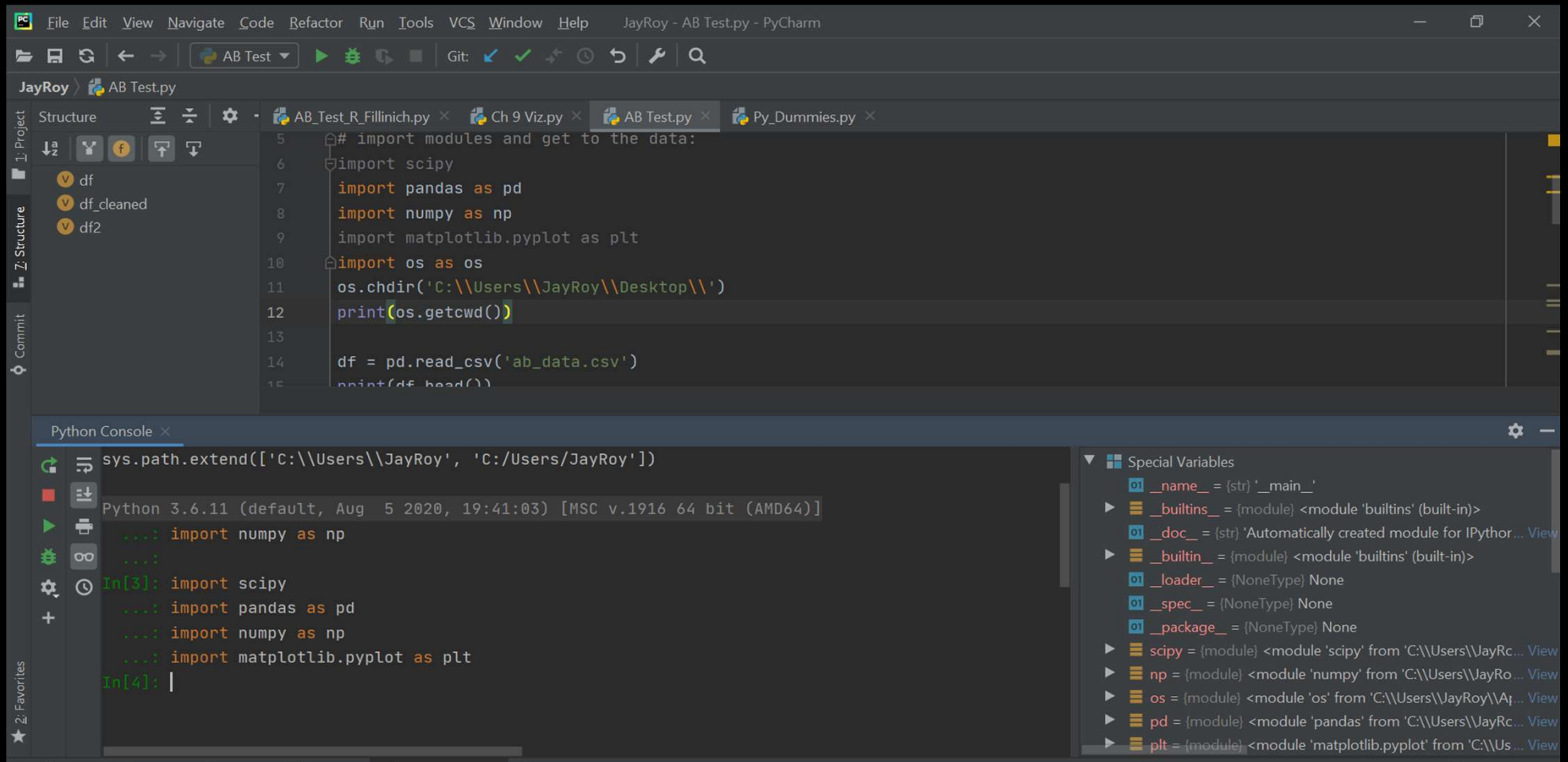
The current cell is active, and its input is:

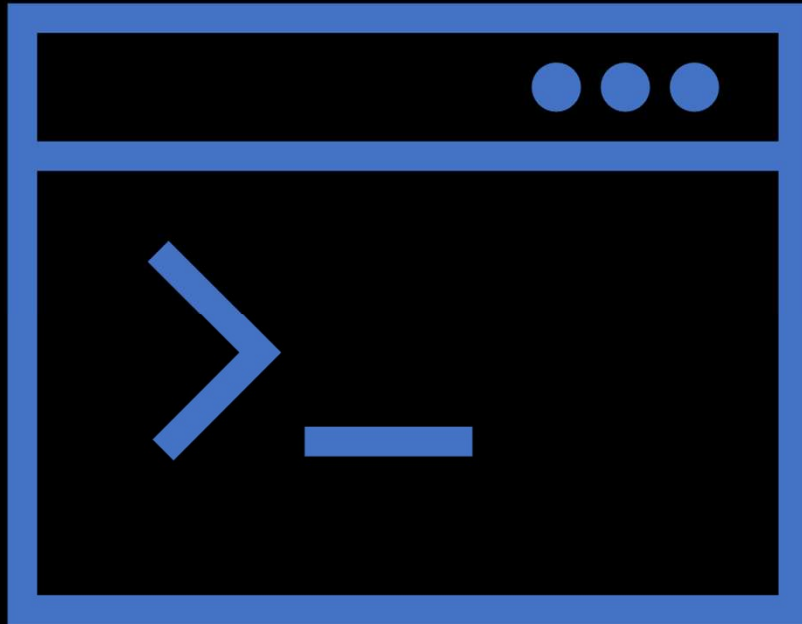
```
In [ ]: housing_df = pd.read_csv('WestRoxbury.csv')
```

Below this, there are three more input cells that have not yet been executed:

```
In [ ]: housing_df.shape
In [ ]: housing_df.head()
In [ ]: housing_df.columns
```

Example of Running Pycharm IDE





What Tool to Start With?

- What type of analyses interest you?
- Ease of Use.
- Ramp up time.
- Computer Literacy knowledge.
- Industry / Location.
- No perfect solution, just begin.



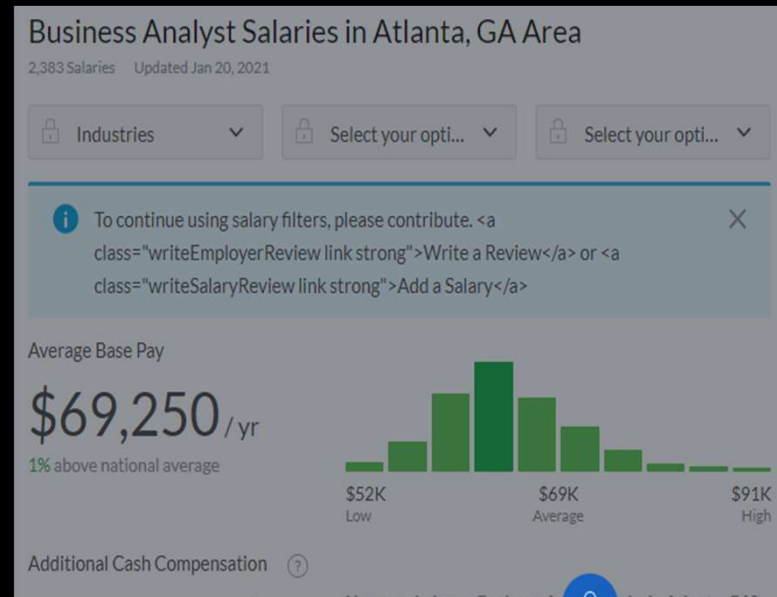
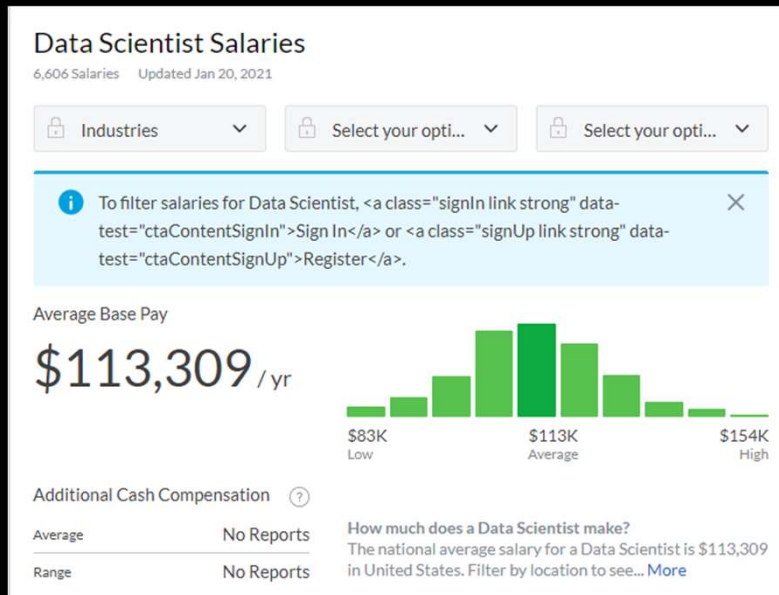
Transformation for BAs and Practical Strategies

Why Transform to “ Data Side ” ?

1. Data is the “ New Oil ”
2. Job and Wage Compression.
3. World is Changing Too Fast. Don't wait!
4. Job Growth and Opportunities Abound.
5. Best Time Ever! Really!



Did You Know ... Data Scientist Salary is ?



Difference of \$ 43,000. – You could buy a couple pairs of shoes / nice car!

How to Transform ?

Suggestions ...

1. Deal with your “Challenges” or “Excuses”.
2. Figure out how you best learn.
3. Ask Lots of Questions. Some even Silly. 😊
4. Partner with others @ your level or higher.
5. Choose – “Start Small” or “Go Big”.



Data Science is not for everyone,
but you can find a place given your talents.



Lessons Learned and Revisited

When Learning
Good Days &
Bad Days.

Technical skills are
developed with
practice, consistency
and initiative.

Deadlines are great
motivators, use
them to your
advantage.

Success is not
binary, it's a
continuum.





ENVISION THE AREA
OF YOUR EXPERTISE



LEARN BY DOING



START DABBLING BY
INSTALL AN IDE



USE EXCEL DATA
SCIENCE FUNCTIONS.



BRUSH UP ON STATS
PROBABILITIES



PERFORM BASIC SQL
QUERIES



FAMILIAR WITH
STATISTICAL
VISUALIZATIONS

Where do I Start ?

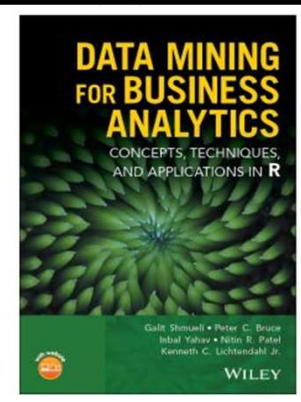
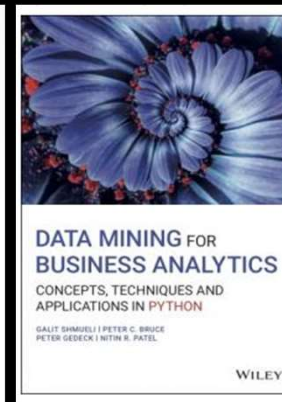
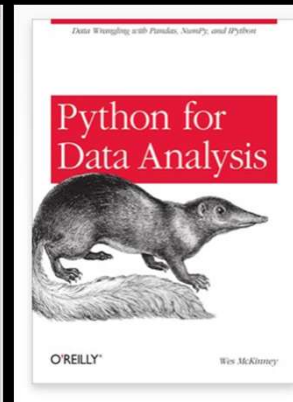
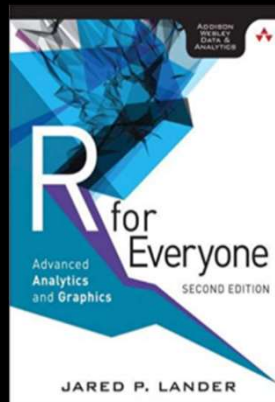
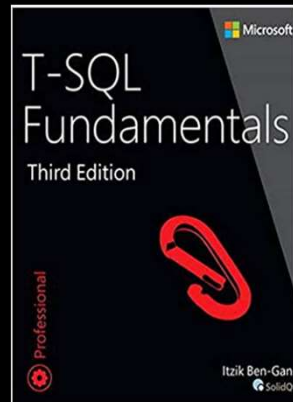
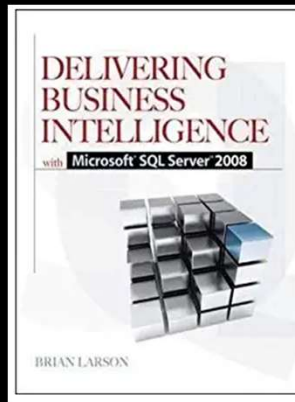
Types of Resources to Help You Transform

- Books
- Variety of MOOCs
- Tutorials / Blogs
- Conferences / Meetup / User Groups
- Ask your developers at your company
- College Continuing Ed classes.
- Coding Bootcamps
- Github – search author's repository for code / examples.
- Search government sites like data.gov

Resources used in my Journey

 Medium

 stackoverflow



Jay Roy, CPA, MBA

jayroyhealthcareitpro@gmail.com

<https://www.linkedin.com/in/jayroydashboards1/>

M: 931-919-8767

Github:

<https://github.com/jayroy1/Beyond-Analysis-Conference-ATL>



- CIO / CFO / CPO
- Product Data Scientist
- Health Tech Entrepreneur
- Educator / Mentor

