f (https://www.facebook.com/AnalyticsVidhya)

8+ (https://plus.google.com/+Analyticsvidhya/posts)

in (https://www.linkedin.com/groups/Analytics-Vidhya-Learn-everything-about-5057165)







(http://datahack.analyticsvidhya.com/contest/date-

your-data)

Home (http://www.analyticsvidhya.com/) > Big data (http://www.analyticsvidhya.com/blog/category/big-data/) > Diag...

# Diagnosing residual plots in linear regression models

BIG DATA (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BIG-DATA/) BUSINESS ANALYTICS

(HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/) SAS

(HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/SAS/)

 $\textbf{f} \ (http://www.facebook.com/sharer.php?u=http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-plots-re$ 

&t=Diagnosing%20residual%20plots%20in%20linear%20regression%20models) ♥ (https://twitter.com/home?

Diagnosing%20residual%20plots%20in%20linear%20regression%20models+http://www.analyticsvidhya.com/blog/2013/12/residual-egression-model/) 8+ (https://plus.google.com/share?url=http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-

) **P** (http://pinterest.com/pin/create/button/?url=http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-&media=http://www.analyticsvidhya.com/wp-

t/uploads/2013/12/Diagnosis.jpg&description=Diagnosing%20residual%20plots%20in%20linear%20regression%20models)





Global Research

(http://admissions.bridgesom.com/pba-new/? utm\_source=AV&utm\_medium=Banner&utm\_campaign=AVBanner)

## My first analytics project involved predicting business from each sales agent and coming up with a targeted intervention for each agent.

I built my first linear regression model after devoting a good amount of time on data cleaning and variable preparation. Now was the time to access the predictive power of the model. I got a MAPE of 5%, Gini coefficient of 82% and a high R-square. Gini and MAPE are metrics to gauge the predictive power of linear regression model. Such Gini coefficient and MAPE for an insurance industry sales prediction are considered to be way better than average. To validate the overall prediction we found the aggregate business in an out of time sample. I was shocked to see that the total expected business was not even 80% of the actual business. With such high lift and concordant ratio, I failed to understand what was going wrong. I decided to read more on statistical details of the model. With a better understanding of the model, I started analyzing the model on different dimensions. After a close examination of residual plots, I found that one of the predictor variables had a square relationship with the output variable.



(http://i2.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/12/Diagnosis.jpg)

Since then, I validate all the assumptions of the model even before reading the predictive power of the model. This article will take you through all the assumptions in a linear regression and how to validate assumptions and diagnose relationship using residual plots.

### **Assumptions of Linear Regression Model:**

There are number of assumptions of a linear regression model. In modeling, we normally check for five of the assumptions. These are as follows:

- 1. Relationship between the outcomes and the predictors is linear.
- 2. Error term has mean almost equal to zero for each value of outcome.
- 3. Error term has constant variance.

- 4. Errors are uncorrelated.
- 5. Errors are normally distributed or we have an adequate sample size to rely on large sample theory.

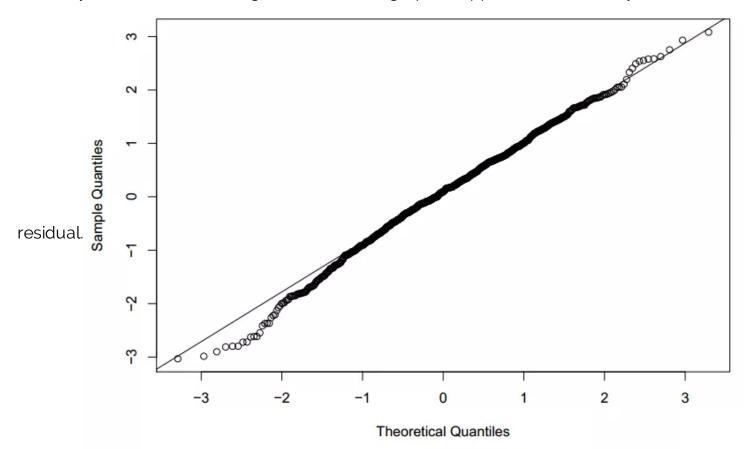
The point to be noted here is that none of these assumptions can be validated by R-square chart, F-statistics or any other model accuracy plots. On the other hand, if any of the assumptions are violated, chances are high that accuracy plot can give misleading results.

#### How to use residual for diagnostics:

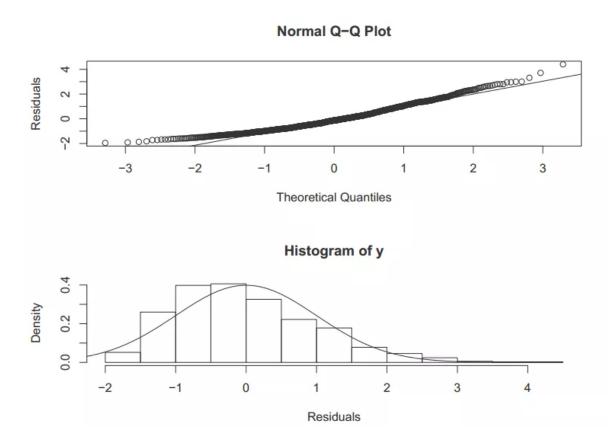
Residual analysis is usually done graphically. Following are the two category of graphs we normally look at:

**1. Quantile plots**: This type of is to assess whether the distribution of the residual is normal or not. The graph is between the actual distribution of residual quantiles and a perfectly normal distribution residuals. If the graph is perfectly overlaying on the diagonal, the residual is

normally distributed. Following is an illustrative graph of approximate normally distributed



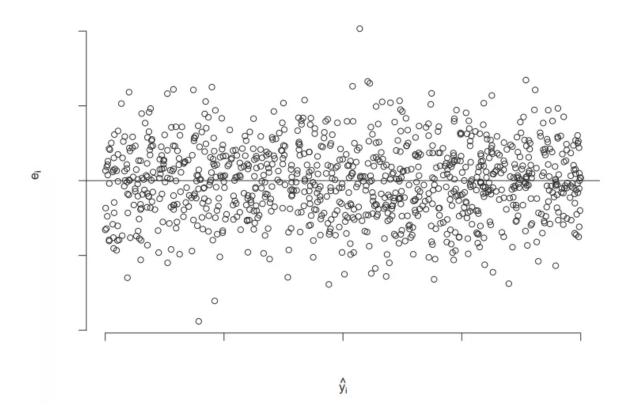
Let's try to visualize a quantile plot of a biased residual distribution.



(http://io.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/11/skewed1.png)

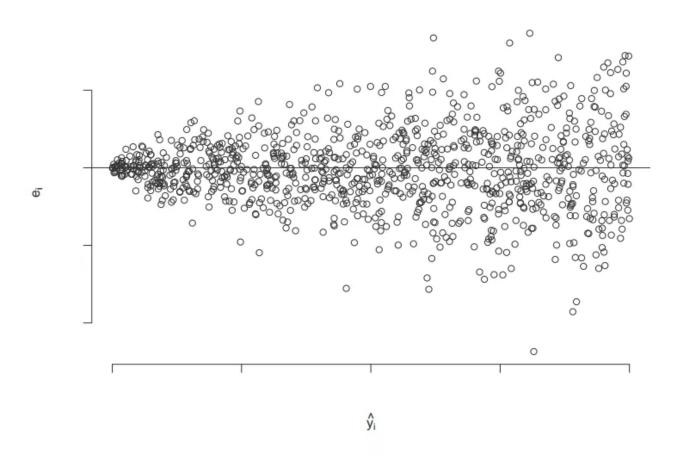
In the graph above, we see the assumption of the residual normal distribution being clearly violated.

**2. Scatter plots:** This type of graph is used to assess model assumptions, such as constant variance and linearity, and to identify potential outliers. Following is a scatter plot of perfect residual distribution



(http://io.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/11/Scatter\_plot.png)

Let's try to visualize a scatter plot of residual distribution which has unequal variance.



(http://io.wp.com/www.analyticsvidhya.com/blog/wpcontent/uploads/2013/11/Scatter\_plot\_2.png)

In the graph above, we see the assumption of the residual normal distribution being clearly violated.





le of single variable regression model to analyze residual llowed for multi-variable as well.

or and the output variable is as follows:

 $0.25x_{i}^{2}$ 

(nttp://www.anatyticsvignya.com/blog/2010/02/quick-(nttp://i2.wp.com/www.anatyticsvignya.com/blog/wp-content/uploads/2013/11/Actualinsights-anatytics-big-data-salaryeg.png) report 2016/)

Ignorant of the type of relationship, we start the analysis with the following equation.

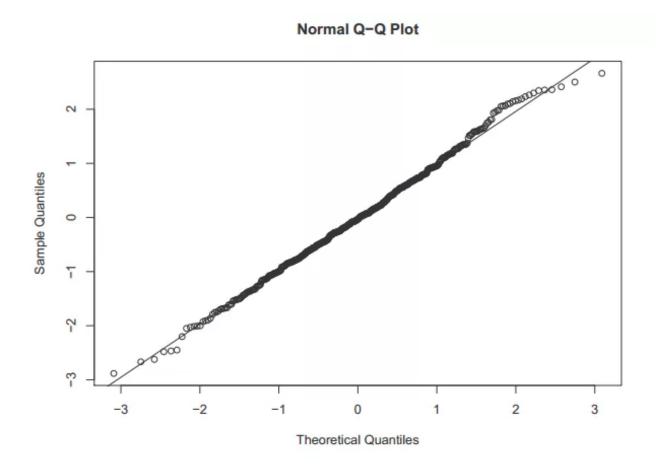
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

(http://i2.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/11/Example2.png)

Can we diagnose this misfit using residual curves?

After making a comprehensive model, we check all the diagnostic curves. Following is the Q-Q plot for the residual of the final linear equation.

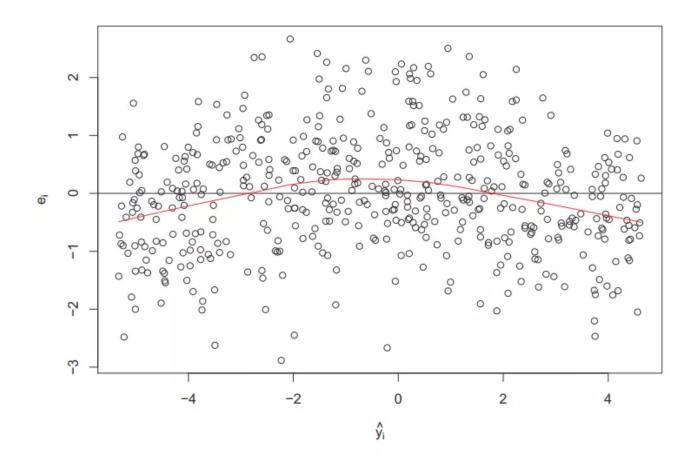
http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/



(http://io.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/11/Example3.png)

Q-Q plot looks slightly deviated from the baseline, but on both the sides of the baseline. This indicated residuals are distributed approximately in a normal fashion.

Following is the scatter plot of the residual:



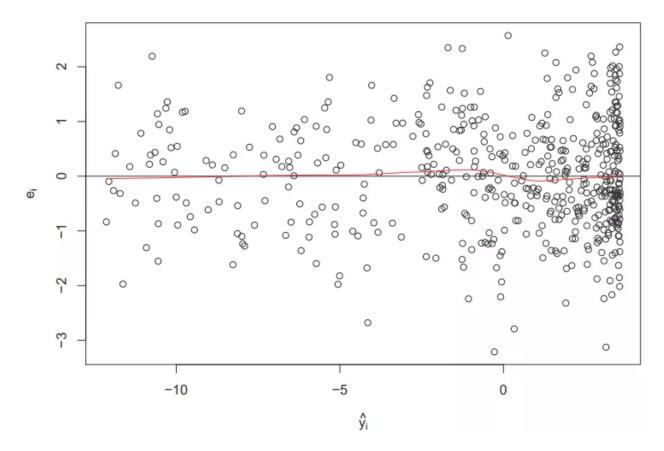
(http://io.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/11/Example4.png)

Clearly, we see the mean of residual not restricting its value at zero. We also see a parabolic trend of the residual mean. This indicates the predictor variable is also present in squared form. Now, let's modify the initial equation to the following equation:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

(http://io.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/11/Example5.png)

Following is the new scatter plot for the residual of the new equation :



(http://io.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/11/Example6.png)

We now clearly see a random distribution and a approximate zero residual mean.

#### **End Notes:**

Every linear regression model should be validated on all the residual plots. Such regression plots directionally guides us to the right form of equations to start with. You might also be interested in the previous article on regression (http://www.analyticsvidhya.com/blog/2013/10/trick-enhance-power-regression-model-2/(http://www.analyticsvidhya.com/blog/2013/10/trick-enhance-power-regression-model-2/))

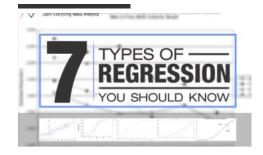
Do you think this provides a solution to any problem you face? Are there any other techniques you use to detect the right form of relationship between predictor and output variables? Do let us know your thoughts in the comments below.

If you like what you just read & want to continue your analytics learning, subscribe to our emails (http://feedburner.google.com/fb/a/mailverify? uri=analyticsvidhya) or like our facebook page (http://facebook.com/analyticsvidhya).

#### **Share this:**

- in (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?share=linkedin&nb=1)
- f (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?share=facebook&nb=1)
- G+ (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?share=google-plus-1&nb=1)
- (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?share=twitter&nb=1)
- (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?share=pocket&nb=1)
- (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?share=reddit&nb=1)

## **RELATED**



(http://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/)



(http://www.analyticsvidhya.com/blog/2015/02/avoid-over-fitting-regularization/)



(http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/)

7 Types of Regression Techniques vou should know! (http://www.analyticsvidhya.com /blog/2015/08/comprehensiveguide-regression/)

In "Business Analytics"

How to avoid Over-fitting using **Regularization?** (http://www.analyticsvidhya.com /blog/2015/02/avoid-over-fittingregularization/)

In "Business Analytics"

**Guide to Build Better Predictive Models using Segmentation** (http://www.analyticsvidhya.com /blog/2016/02/guide-buildpredictive-models-segmentation/)

In "Business Analytics"

TAGS: BUSINESS ANALYTICS (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BUSINESS-ANALYTICS/), LINEAR-REGRESSION (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/LINEAR-REGRESSION/). PREDICTIVE MODELING (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/PREDICTIVE-MODELING/), REGRESSION (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/REGRESSION/), RESIDUAL (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/RESIDUAL/), RESIDUAL PLOTS (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/RESIDUAL-PLOTS/). SAS (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/SAS/). VALIDATION (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/VALIDATION/)

**Previous Article** Interview with data scientist and top Kaggler, Mr. Steve Donoho top-data-scientist-kaggler-mr-steve-

donoho/)

**Next Article** How do banks identify the next best product need of its customer? (http://www.analyticsvidhya.com/blog/2013/11/intervilentp://www.analyticsvidhya.com/blog/2013/12/banksidentify-product-customer/)



(http://www.analyticsvidhya.com/blog/author/tavish1/) **Author** 

# Tavish Srivastava (http://www.analyticsvidhya.com/blog/author/tavish1/)

I am Tavish Srivastava, a post graduate from IIT Madras in Mechanical Engineering. I have more than two years of work experience in Analytics. My experience ranges from hands on analytics in a developing country like India to convince banking partners with analytical solution in matured market like US. For last two and a half years I have contributed to various sales strategies, marketing strategies and Recruitment strategies in both Insurance and Banking industry.

## 9 COMMENTS



**Tuhin Chartenpadinyangang**Eyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?replytocom=3482#respond) december 2, 2013 at 11:37 am (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/#comment-3482)

Good one Kunal, though I prefer to use designated tests to verify each of the assumptions.



Tavish Srimastava sama. Analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?replytocom=3486#respond)
December 2, 2013 at 11:45 am (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regressionModel/#comment-3486)

Tuhin,

I agree doing designated test to verify each assumption is a more sturdy approach. However, using the approach mentioned in the article will tell you two facts. First, whether assumption is validated or not. Second, if assumptions are not validated, what will be the best starting point for the next iteration. For instance, in the example mentioned in the article, residual plot gave us a better starting equation. Please share your thoughts of how will designated tests help on the second point.

Thanks, Tavish



Kunal Jain (http://www.analyticsvidlum/600)2053 Latesidual-plots-regression-model/?replytocom=3521#respond)
December 2, 2013 at 4:38 PM (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regressionModel/#comment-3521)

Tuhin.

While the designated tests would provide you the output, they do not provide as much details about why things have failed.

Also, even if you use designated tests, it makes sense to understand what has failed.

Thanks,

Kunal



**Tuhin Chateprathyawsaws**!yticsvidhya.com/blog/2013/12/residual-plots-regression-model/?replytocom=3622#respond) december 3, 2013 at 6:09 am (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/#comment-3622)

Dear Kunal and Tavish.

Both the presumptive and confirmatory tests are required in any field of science. It's not about a fight of graphical method vs. the tests. It's about performing both the things sequentially. For instance, you can have a look of the distribution (normal) graphically at first by histogram of residuals or normal probability plot (NPP) and then apply Anderson Darling test in Minitab or Jarque Bera test in E-Views or Kolmogorov-Smirnov test and Shapiro Wilk test in SPSS to confirm. That's how I perform it.



Kunal Jain (http://www.analyticsvid.hva.com/269/269/268/28) Residual-Plots-Regression-Model/?Replytocom=3771#Respond) December 5, 2013 at 5:04 pm (http://www.analyticsvidhya.com/blog/2013/12/Residual-Plots-Regression-Model/#Comment-3771)

Completely agree Tuhin with your approach.

Regards,

Kunal



Rakesh Jap saysp://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?replytocom=85289#respond)
April 28, 2015 at 12:50 pm (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regressionmodel/#comment-85289)

Is not there any standard set of designated tests which one can use for confirmation?



**Tuhin Chattapathyawaawse**yticsvidhya.com/blog/2013/12/residual-plots-regression-model/?replytocom=3483#respond)
DECEMBER 2, 2013 AT 11:39 AM (HTTP://www.analyticsvidhya.com/blog/2013/12/residual-plots-regressionMODEL/#COMMENT-3483)

Sorry, it's by Tavish. Liked it Tavish.



## Deepak Rai says:



## DECEMBER 13RE2013HATP4/2NAWAMANAUTTPCS/WOWVALGONALDIOG/2013/12/TRESIONA/BRIOG/SETEISPESS/FORISM DDEAL/PREDINS/OCCORRES/SEMERESPOND) MODEL/#COMMENT-3879)

Hi,

How to we check whether Error Terms are independent to each other i.e. they are not correlated with one another (Assumption of Independence of the Error Terms) and Colinearity of the independent variables?

Regards,

Deepak Rai



Tavish Sax of (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/?replytocom=64003#respond)
January 9, 2015 at 9:56 am (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regressionModel/#Comment-64003)

Deepak,

This will be visible from your residual scatter plot with independent variables. You will see that your error will show you autocorrelation i.e. you will that a positive error tend to follow a positive error and vice versa.

Hope this helps.

Tavish

## **LEAVE A REPLY**

Connect with:



(http://www.analyticsvidhya.com/wp-login.php?

action=wordpress\_social\_authenticate&mode=login&provider=Facebook&redirect\_to=http%3A%2F%2F\plots-regression-model%2F)

Your email address will not be published.

Comment		
		<i>h</i>

Name (required)

Email (required)

Website

SUBMIT COMMENT

- Notify me of follow-up comments by email.
- Notify me of new posts by email.

## **TOP USERS**

Rank	Name		Points
1		Nalin Pasricha (http://datahack.analyticsvidhya.com/user/profile/Nalin)	3478
2		SRK (http://datahack.analyticsvidhya.com/user/profile/SRK)	3364
3		Aayushmnit (http://datahack.analyticsvidhya.com/user/profile/aayushmnit)	3075



binga (http://datahack.analyticsvidhya.com/user/profile/binga) 2623

vikash (http://datahack.analyticsvidhya.com/user/profile/vikash) 2190

More Rankings (http://datahack.analyticsvidhya.com/users)



(http://pgpba.greatlakes.edu.in/?

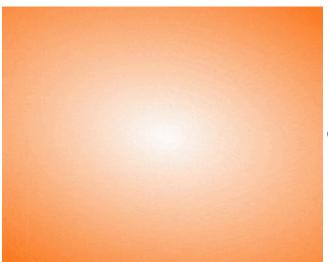
utm\_source=AVM&utm\_medium=Banner&utm\_campaign=Pgpba\_decjan)

## **POPULAR POSTS**

- Free Must Read Books on Statistics & Mathematics for Data Science
   (http://www.analyticsvidhya.com/blog/2016/02/free-read-books-statistics-mathematics-data-science/)
- A Complete Tutorial to Learn Data Science with Python from Scratch
   (http://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/)
- Essentials of Machine Learning Algorithms (with Python and R Codes)
   (http://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/)

- A Complete Tutorial on Time Series Modeling in R
   (http://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/)
- Complete guide to create a Time Series Forecast (with Codes in Python)
   (http://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/)
- 4 tricky SAS questions commonly asked in interview
   (http://www.analyticsvidhya.com/blog/2013/11/4-sas-tricky-analytics-interview/)
- SAS vs. R (vs. Python) which tool should I learn?(http://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/)
- 7 Important Model Evaluation Error Metrics Everyone should know
   (http://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/)





(http://imarticus.org/programs/business-analytics-

professional/)

## **RECENT POSTS**



(http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-

science-scratch/)

A Complete Tutorial to learn Data Science in R from Scratch (http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/)
MANISH SARASWAT, FEBRUARY 28, 2016



(http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-

segmentation/)

Guide to Build Better Predictive Models using Segmentation (http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/) GUEST BLOG, FEBRUARY 26, 2016



(http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-

salary-report-2016/)

Quick Insights: India Analytics and Big Data Salary Report 2016 (http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/)

KUNAL JAIN, FEBRUARY 24, 2016



(http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-

India Exclusive: Analytics and Big Data Salary Report 2016 (http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/) KUNAL JAIN . FEBRUARY 22, 2016



(http://www.edvancer.in/certified-business-analytics?

utm\_source=AV&utm\_medium=AVads&utm\_campaign=AVads1&utm\_content=cbapavad)

## **GET CONNECTED**



4,162

**FOLLOWERS** 

(http://www.twitter.com/analyticsvidhya)

**8**+

915

**FOLLOWERS** 

(https://plus.google.com/+Analyticsvidhya)

f 11,954

FOLLOWERS

(http://www.facebook.com/Analyticsvidhya)

∌ Fmai

**SUBSCRIBE** 

(http://feedburner.google.com/fb/a/mailverify? uri=analyticsvidhya)

## **ABOUT US**

For those of you, who are wondering what is "Analytics Vidhya", "Analytics" can be defined as the science of extracting insights from raw data. The spectrum of analytics starts from capturing data and evolves into using insights / trends from this data to make informed decisions.

## **STAY CONNECTED**



4,162

**FOLLOWERS** 

(http://www.twitter.com/analyticsvidhya)



915

**FOLLOWERS** 

(https://plus.google.com/+Analyticsvidhya)

### LATEST POSTS



**FOLLOWERS** 

(http://www.facebook.com/Analyticsvidhya)



SUBSCRIBE

(http://feedburner.google.com/fb/a/mailverify? uri=analyticsvidhya)



(http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-

science-scratch/)

A Complete Tutorial to learn Data Science in R from Scratch (http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/)
MANISH SARASWAT, FEBRUARY 28, 2016



(http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-

segmentation/)

Guide to Build Better Predictive Models using Segmentation (http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/)
GUEST BLOG . FEBRUARY 26, 2016



(http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-

salary-report-2016/)

Quick Insights: India Analytics and Big Data Salary Report 2016 (http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/)

KUNAL JAIN, FEBRUARY 24, 2016

India Exclusive: Analytics and Big Data Salary Report 2016



(http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-

## (http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/) KUNAL JAIN , FEBRUARY 22, 2016

## **QUICK LINKS**

Home (http://www.analyticsvidhya.com/)
About Us (http://www.analyticsvidhya.com/about-me/)
Our team (http://www.analyticsvidhya.com/about-me/team/)

Privacy Policy

(http://www.analyticsvidhya.com/privacy-policy/)

Refund Policy

(http://www.analyticsvidhya.com/refund-policy/)

Terms of Use

(http://www.analyticsvidhya.com/terms/)

## **TOP REVIEWS**

© Copyright 2015 Analytics Vidhya