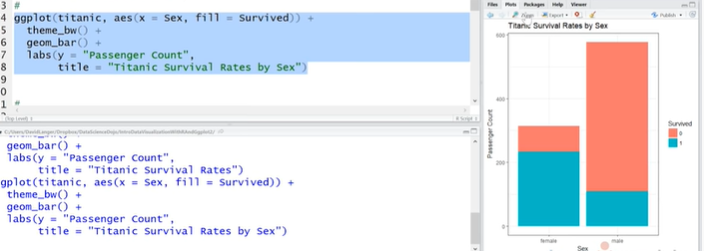
**Module 2**

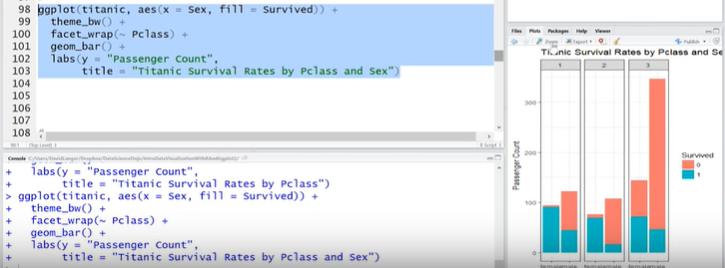
* Problem Definition Framework
  + Define
    - What hypothesis
  + Evaluate
    - Do we have everything we need
  + Prioritize
  + Prepare
* Effective Problem Definition
  + Clarity on business issue
  + Testable hypothesis
  + Ability to assess outcome through clear KPIs
* Evaluate and Prioritize
  + Do you have the data
  + Is it sufficient quality
  + Other considerations
    - Regulatory acceptance
    - Customer acceptance
    - Potential for gaming

**Module 3**

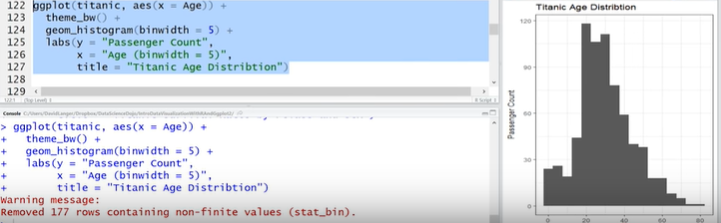
* Use ‘fill’ in ggplot with categorical variable to separate bar by colours



* Use facet\_wrap to layer 2 categorical variables together



* For histograms ‘bin\_width’ is essentially the qcut function



* Density plots are useful:



**Module 6**

* Model Diagnostics
  + Use drop1 function to test AIC scores when dropping each variable
* Regularization
  + Restricting the “magnitude” of the inputs with a penalty term for more/larger parameters
  + By adding the regularization term, the model will only include the parameter when it is worth its penalty
  + Inputs should be standardized if using this approach
* Types of Regularized Regression
  + Ride Regression
    - Uses the sum of squared parameters as the error term
  + Lasso regression
    - Uses mean absolute error as penalty term
    - Has ability to completely remove a feature, whereas ridge will penalize parameters only to the point they become very small
  + Elastic net regression
    - Uses combination of both, with an alpha term where the weight is alpha for one and 1 – alpha for the other
* Limitations of Regularization for feature selection
  + Feature selection becomes automatic and is not always interpretable
  + It is model dependent, so if we choose linear regression that’s different than other models