5 章　探索的データ分析

EDA（Exploratory data analysis）

①　データに対する質問　→　②可視化、変換、モデル化　→　③質問の洗練

データに対する質問

間違った質問に対する正確な答えより、正しい質問に対する近似解の方がはるかに優れている（Turkey）

→良質な質問を作り出すために大量の質問を作り出す。

例えば、変数においてどんな変動があるのか？　変数間にはどんな共変動があるのか？

変動の理解（分布の可視化、典型値、異常値）

カテゴリ数の分布 ――― 棒グラフ

連続数の分布 ――― ヒストグラム

【典型値】

●どの値が最も一般的か？　また、それはなぜか？　●複峰型の分布になっているのはなぜか？

●その値が稀か。それはなぜか。それは期待に沿うか。　●0.5 カラットでアイディアルカラットが多いのはなぜか？

●3 カラットより大きいダイヤモンドがないのはなぜか？　など

出現した（プロットした）に対して、いくつの有用な質問を投げかけることができますか？

【異常値（外れ値）】

異常値は入力エラー？　新しい科学的な発見？

※coord_cartesian() で異常値を探す　デカルト座標＋引数で拡大するという意味

外れ値の理由を明らかにするように努めるが、明らかにならないこともある。外れ値の処理方法はドキュメントととして必ず残す。

異常値を欠損値として処理する

共変動の理解

→2 つ以上の変数が関連して同時に変動する傾向

【カテゴリ変数と連続変数】 度数分布多角形、箱ひげ図、バイオリンプロット、geom_jitter

【2 つのカテゴリ変数】 geom_count, geom_tile

【2 つの連続変数】geom_point, geom_bin2d, geom_hex, geom_boxplot(group でカウント)

パターンとモデル

モデルは、データからパターンを抽出する道具。

→Ⅳ部へ