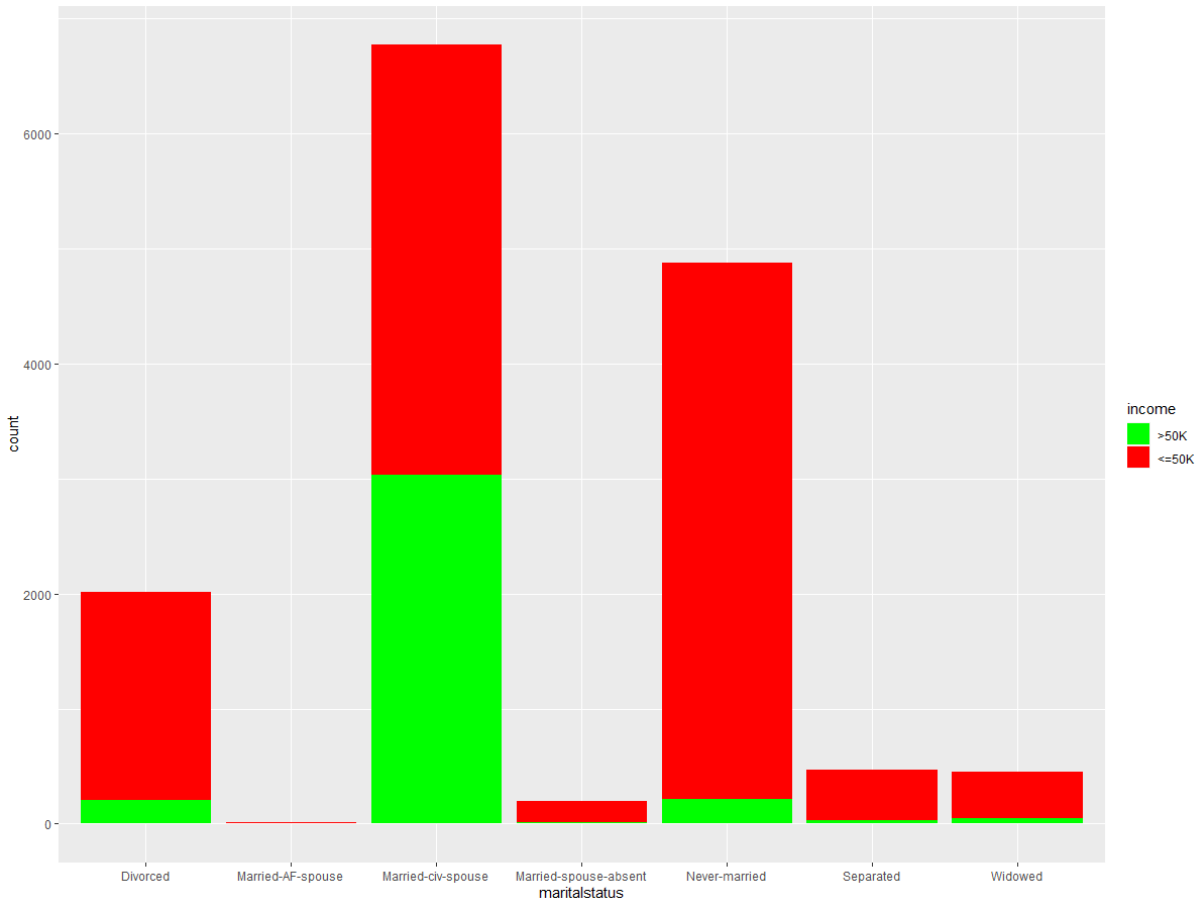
Eksploracja Danych – Raport 1 Klaudia Pełczyńska

Analizowany zbiór danych znajduje się w pliku *Adult_ch3_training.csv*, w którym dostępne są następujące zmienne:

- **age** – wiek klienta
- **workclass** – grupa pracownicza
  - **Federal_gov** – administracja federalna
  - **Local-gov** – administracja lokalna
  - **Never-worked** – nigdy niepracujący
  - **Private** – sektor prywatny
  - **Self-emp-inc** – samo zatrudniony, działalność gospodarcza
  - **Self-emp-not-inc** – samo zatrudniony, bez działalności
  - **State-gov** – administracja stanowa
  - **Without-pay** – bez dochodów
- **education** – lata edukacji
- **maritalstatus** – stan cywilny
  - **Divorced** – rozwiedziony
  - **Married-AF-spouse** – w małżeństwie z osobą z sił zbrojnych
  - **Married-civ-spouse** – w małżeństwie z cywilem
  - **Married-spouse-absent** – w małżeństwie z osobą nieobecną
  - **Never married** – nigdy niezamężna/nieżonaty
  - **Separated** – w separacji
  - **Widowed** – wdowiec/wdowa
- **occupation** – zawód
- **sex** – płeć
  - **Female** – kobieta
  - **Male** – mężczyzna
- **capitalgain** – zysk kapitału
- **capitalloss** – strata kapitału

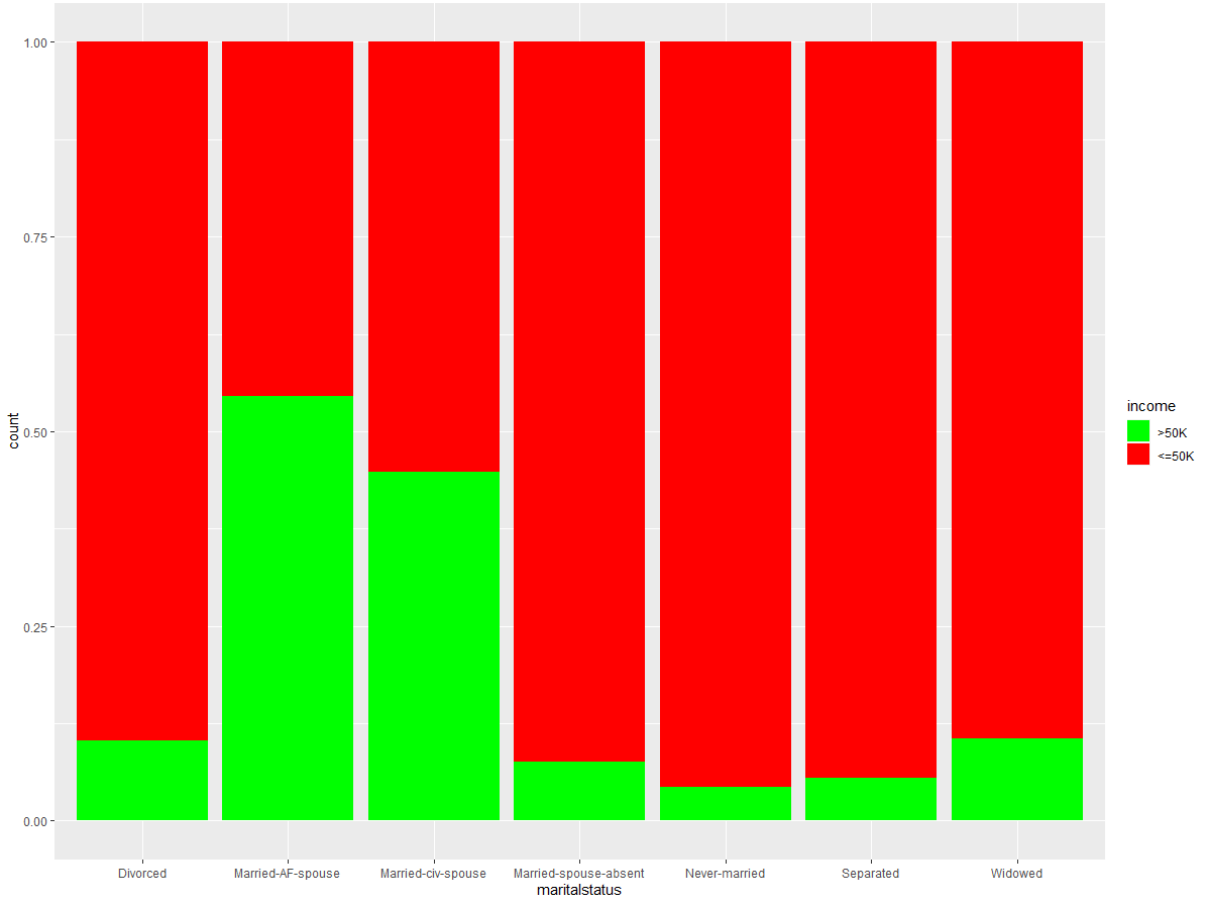
Zmienną celu jest zmienna **income** (dochód) przyjmująca wartości: >50K (ponad 50 tys. dolarów rocznie) oraz <=50K (nie więcej niż 50 tys. rocznie). Dane będą analizowane pod kątem wartości >50K zmiennej income.



Rysunek 1. Zestawiony wykres słupkowy dla zmiennej *maritalstatus* i zmiennej *income*.

Powyższy wykres przedstawia zestawioną liczbę osób podzielonych według ich stanu cywilnego wraz z ich dochodem. Z wykresu możemy odczytać, że najwięcej w naszym zbiorze danych jest osób, które są w małżeństwie z cywilem oraz tych, które nigdy nie były zamężne/żonate. Najmniej natomiast jest osób, które są w małżeństwie z osobą z sił zbrojnych i z osobą nieobecną.

Największa ilość osób, które zarabiają powyżej 50 tys. rocznie znajduje się w grupie osób, które są w małżeństwie z cywilem – jest to około 3000 osób. W pozostałych grupach zdecydowana większość osób zarabia poniżej 50 tys. rocznie.



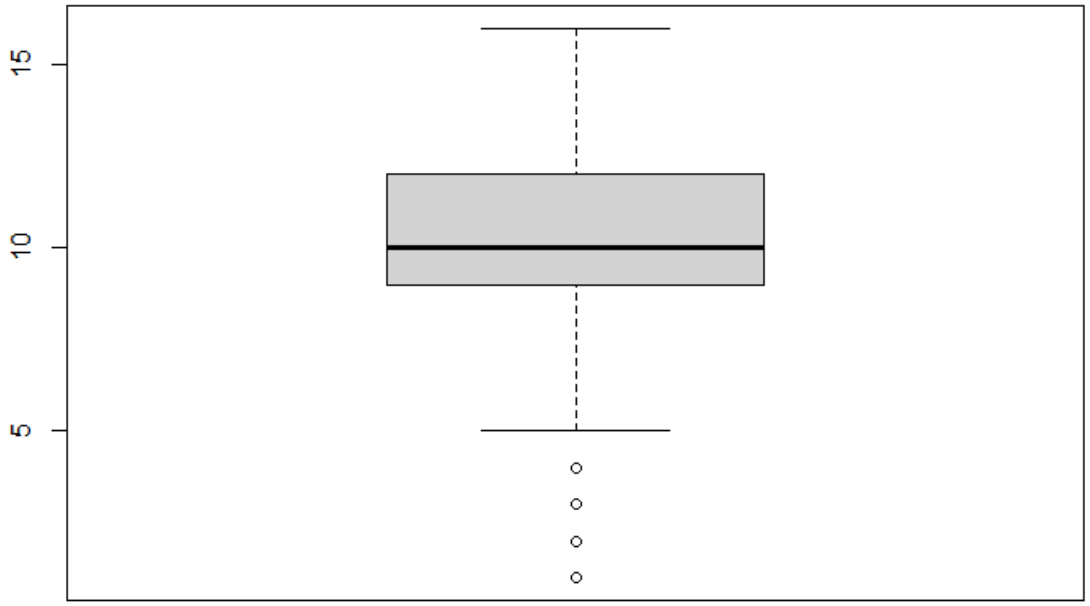
Rysunek 2. Znormalizowany zestawiony wykres słupkowy dla zmiennej *maritalstatus* i zmiennej *income*.

Znormalizowany wykres słupkowy przedstawia stosunek liczby osób zarabiających powyżej 50 tys. rocznie do liczby osób zarabiających poniżej 50 tys. rocznie. Analizując powyższy wykres widzimy, że największy odsetek liczby osób zarabiających powyżej 50 tys. rocznie jest w grupach osób, które są w małżeństwie z osobą z sił zbrojnych oraz tych, które są w małżeństwie z cywilem. W przypadku pierwszej grupy jest to ponad 50%, a w przypadku drugiej grupy około 45%. Natomiast w pozostałych grupach odsetek osób zamożniejszych jest bardzo zbliżony, dla każdej z tych grup jest niższy niż 10%. Najniższy odsetek osób zarabiających powyżej 50 tys. rocznie jest w grupie osób, które nigdy nie były zamężne/żonate i jest on równy około 5%.

	Divorced	***Married-AF-spouse***	***Married-civ-spouse***	***Married-spouse-absent***	***Never married***	***Separated***	***Widowed***
<=50K	89.67%	45.45%	55.15%	92.46%	95.61%	94.49%	89.40%
>50K	10.33%	54.55%	44.85%	7.54%	4.39%	5.51%	10.60%

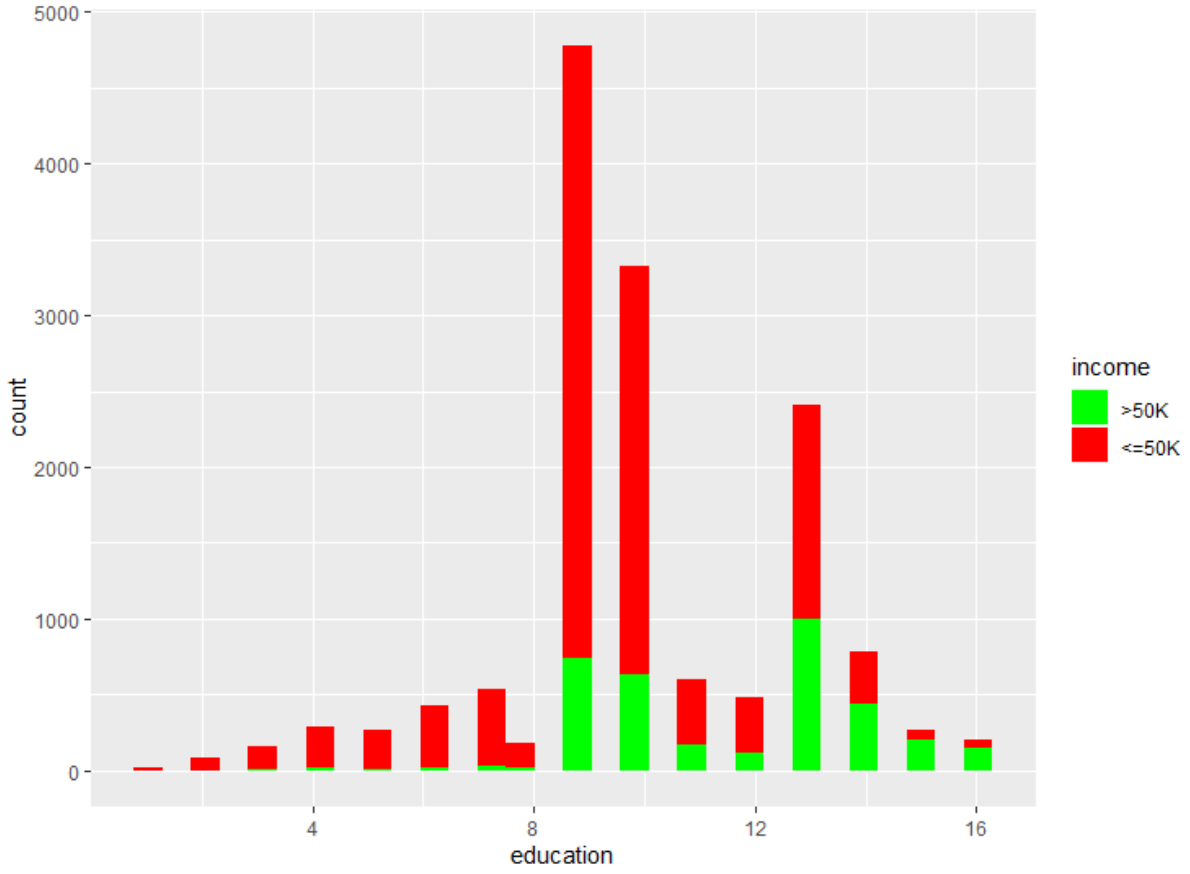
Tabela 1. Tabela krzyżowa dla zmiennej *maritalstatus* i zmiennej *income*.

Uzupełnieniem drugiego wykresu (rysunek 2.) jest powyższa tabela krzyżowa, w której umieszczone są wartości procentowe. Możemy te dane interpretować jako prawdopodobieństwo tego, że osoba o danym statusie cywilnym zarabia powyżej 50 tys. rocznie.



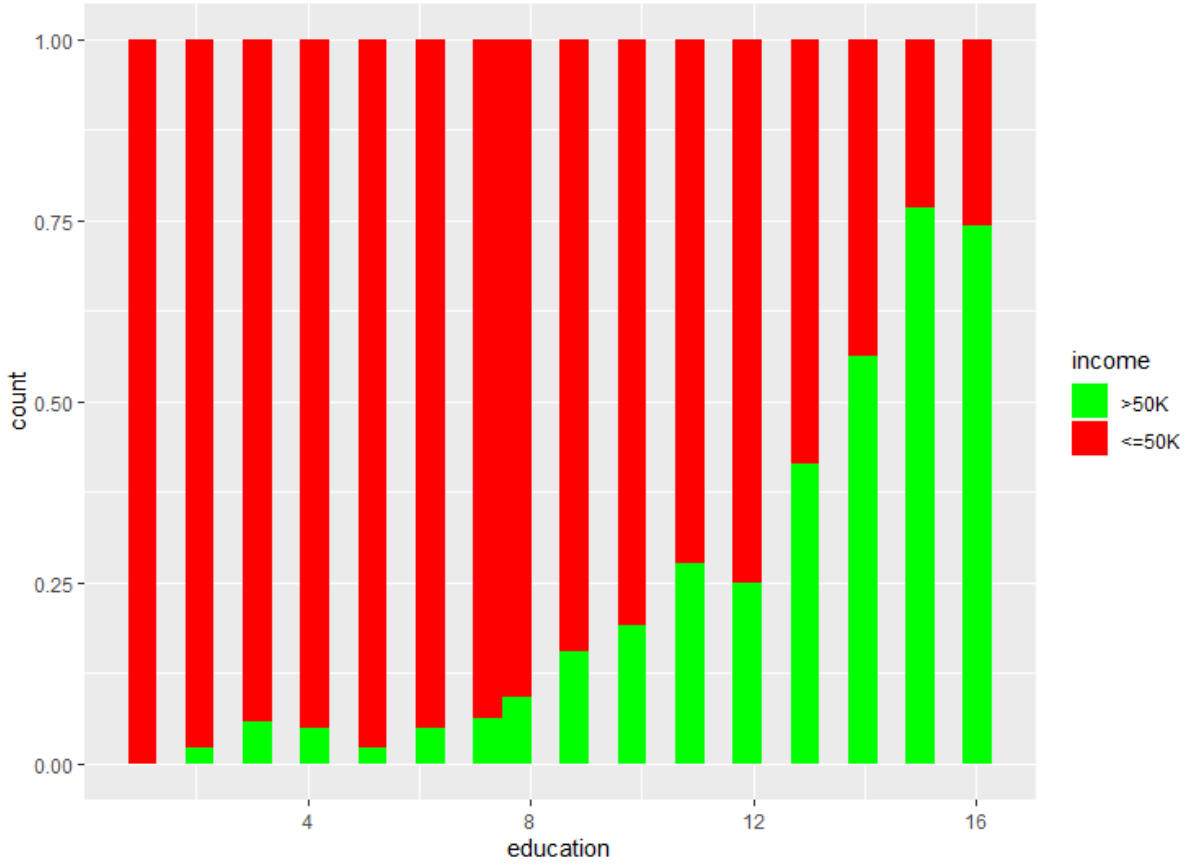
Rysunek 3. Wykres skrzynkowy zmiennej *education*.

Wykres skrzynkowy służy do zobrazowania skośności rozkładu zmiennej education i wartości odstających. Obserwacji odstających jest 113.



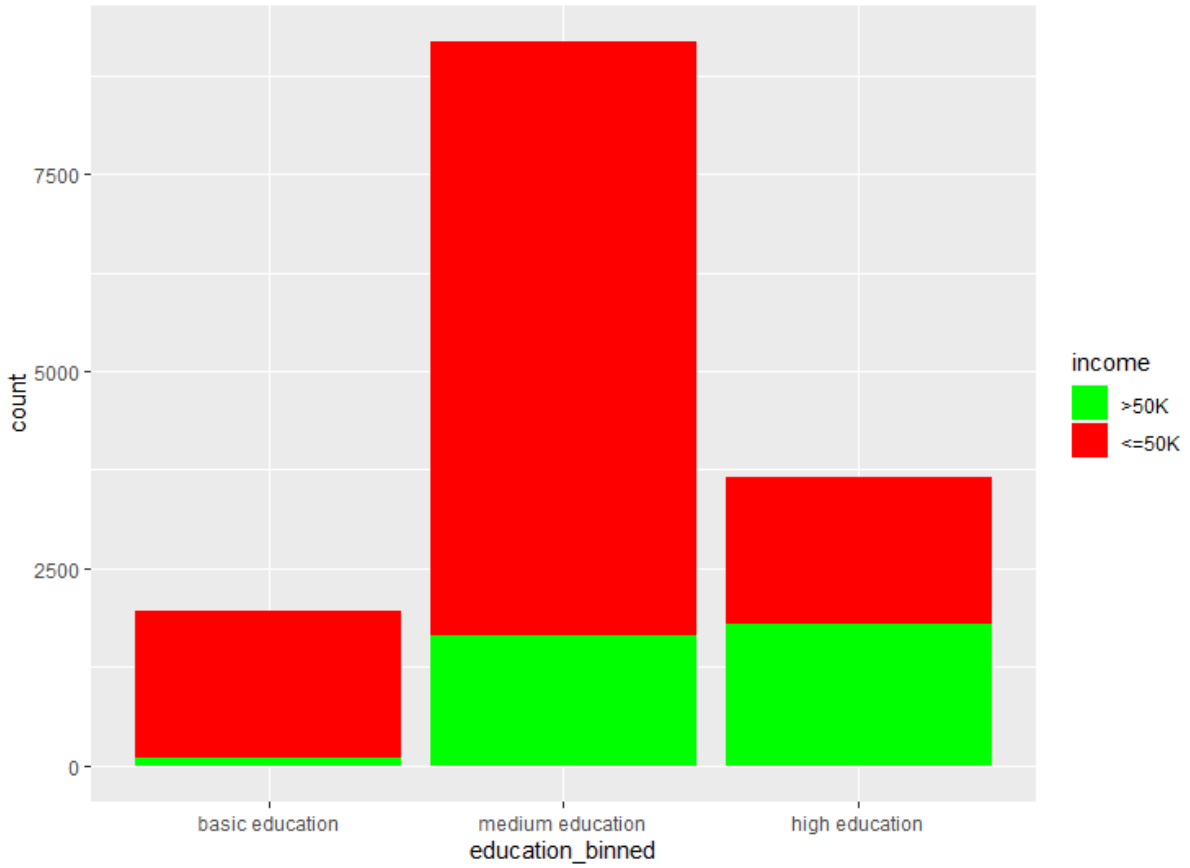
Rysunek 4. Zestawiony histogram dla zmiennej *education* i zmiennej *income*.

Powyższy wykres przedstawia histogram zmiennej education, którego słupki podzielone zostały zmienną income. Możemy zauważyć, że najwięcej osób zakończyło swoją edukację po 9 latach. Najwięcej osób, które zarabiają powyżej 50 tys. rocznie skończyło swoją edukację po 13 latach, natomiast takich osób jest najmniej wśród tych, którzy skończyli swoją edukację po 8 latach lub wcześniej.



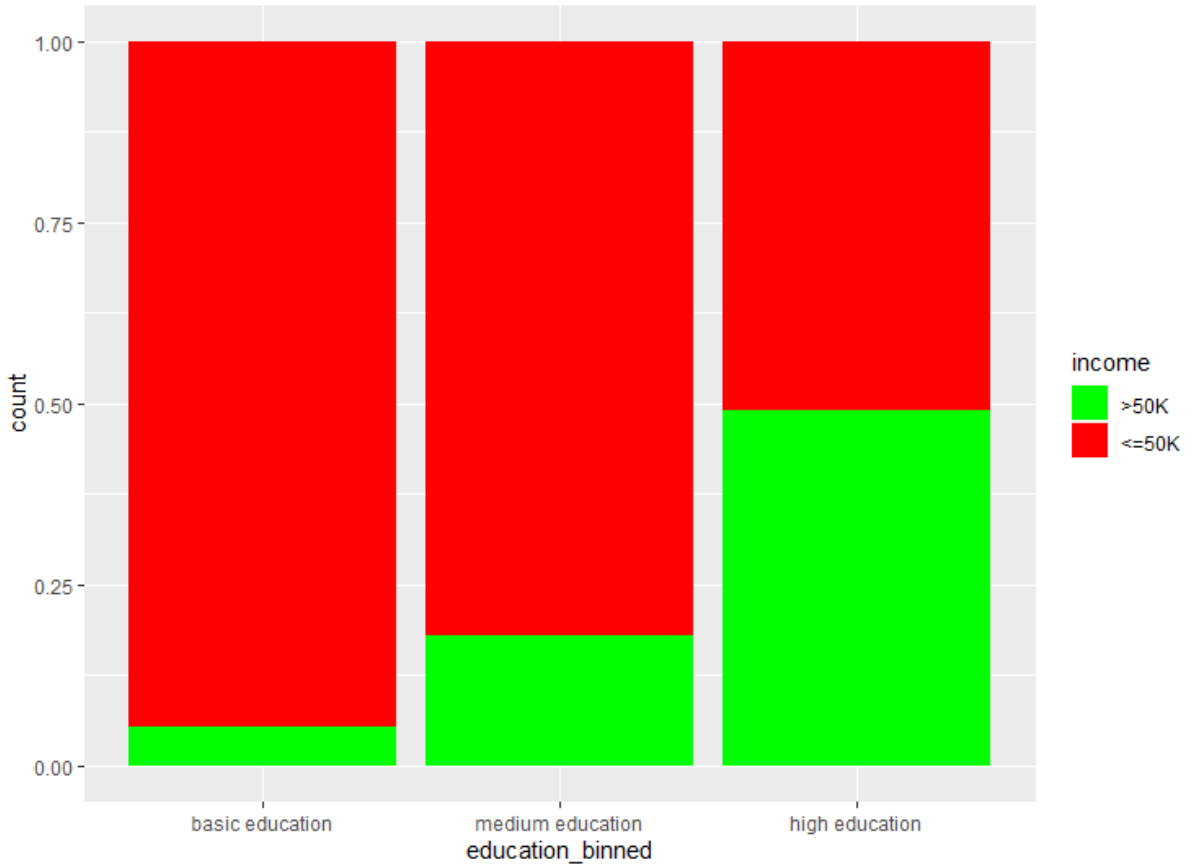
Rysunek 5. Znormalizowany zestawiony histogram dla zmiennej *education* i zmiennej *income*.

Wykres 5 daje nam dużo więcej informacji niż wykres 4. Na wykresie 4 liczba osób zarabiających powyżej 50 tys. rocznie po 16 latach edukacji jest dużo niższa niż wśród osób, które uczyły się 8 lat. Natomiast na powyższym wykresie widać, że odsetek osób zarabiających powyżej 50 tys. rocznie rośnie wraz z wzrostem wartości zmiennej education. Odsetek osób zarabiających powyżej 50 tys. rocznie w każdej z grup możemy interpretować jako prawdopodobieństwo, dlatego można wysnuć następujący wniosek: „Im dłużej będziemy się uczyć tym wyższe będą nasze zarobki.”



Rysunek 6. Zestawiony wykres słupkowy dla zmiennej *education_binned* i zmiennej *income*.

Zestawiony wykres słupkowy dla zmiennej education_binned i zmiennej income przedstawia liczby osób zarabiających powyżej 50 tys. rocznie i zarabiających poniżej 50 tys. rocznie, które zostały pogrupowane w trzy klasy: basic education (education <= 8), medium education (8 < education < 13) i high education (13 <= education). Z powyższego wykresu możemy odczytać, że najwięcej jest osób o wykształceniu średnim, natomiast najmniej o wykształceniu podstawowym.



Rysunek 7. Znormalizowany zestawiony wykres słupkowy dla zmiennej *education_binned* i zmiennej *income*.

Wnioski, które możemy wysnuć z powyższego wykresu są podobne do tych, które zostały sformułowane przy wykresie 5. Na wykresie 6. liczba osób zarabiających powyżej 50 tys. rocznie była zbliżona w grupach: medium i high education. Jednak na wykresie 7 widzimy, że odsetek osób zarabiających powyżej 50 tys. rocznie w grupie high education jest ponad 2 razy większy niż w grupie medium education. Jeśli odsetki zinterpretujemy jako prawdopodobieństwo możemy dojść do wniosków, że prawdopodobieństwo zarobków powyżej 50 tys. rocznie jest wyższe, jeśli nasza edukacja trwała dłużej niż 13 lat.