

**Logiciels statistiques pour l’analyse de données**

**6-613-11**

Devoir 1

Remis par :

William Tankou

Yury Vincent Sambale

Kokou Viwoto

À :

Sarah Legendre Bilodeau

Le 28 Octobre 2016

# Description

Ce projet de nature académique a pour but de nous aider à développer nos compétences techniques dans l’utilisation du langage de programmation R. Le logiciel R studio est utilisé afin de traiter, analyser et présenter des données dans le but de déceler des tendances et d’éventuellement tirer des conclusions. Dans notre cas, un jeu de données comprenant des informations sur les écrasements d’avions qui ont eu lieu entre 1908 et 2009 est utilisé.

# Pourquoi avons-nous choisi ce jeu de données ?

Nos critères principaux pour ce projet étaient d’analyser un jeu de données qui contenait peu de textes, qui ne traitait pas de données économiques et qui traitait d’un sujet original. On entend par original, un sujet qui n’avait pas trop été analysé dans le passé et dont peu de conclusions ont été tirées. En faisant plusieurs recherches, nous sommes tombés sur le site web [www.kaggle.com](http://www.kaggle.com) qui contenait plusieurs jeux de données intéressants sur des sujets diverses. Le jeu de données sur les écrasements d’avions nous a paru intéressant puisqu’il répondait à tous nos critères en plus d’avoir beaucoup de détails sur les écrasements.

# Description du jeu de données

Le jeu de données couvre 5268 crashs d’avions qui ont eu lieu entre 1908 et 2009. Les 13 colonnes contiennent l’information suivante :

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Date | Date du crash |
| Time | Heure du crash |
| Location | Endroit du crash |
| Operator | Opérateur de l'avion |
| Flight | Numéro de vol |
| Route | Itinéraire prévu |
| Type | Modèle d’avion |
| Registration | Numéro de registration de l’avion |
| Cn.In | Aucune description fournie |
| Aboard | Nombre de passagers à bord |
| Fatalities | Nombre de morts |
| Ground | Aucune description fournie |
| Summary | Détails du crash. Les données sous forme de texte |

# Données Manquantes

Il est important de noter que notre jeu de données possède quelques lacunes:

1. Pour des raisons inconnues, notre fichier de données n’a enregistré aucun crash entre 1908 et 1912, mis à part un crash en 1908.
2. Les informations sur les crashs survenus au cours des conflits armés semblent aussi manquantes. En effet, on s’attendrait à voir une hausse beaucoup plus importante du nombre d’écrasement militaires au cours de la période allant de 1939 à 1945 (seconde guerre mondiale). Cependant le jeu de données nous produit en moyenne 50 crashs par an au cours de cette période.
3. Certaines variables sont manquantes pour plusieurs des crashs répertoriés dans le jeu de données.
4. Plusieurs variables ont des erreurs d’orthographe, dû au fait que la saisie des données a été faites manuellement
5. Les noms de pays étaient mal écrits dans beaucoup de situation. Il fallait donc normaliser et corriger ces erreurs avant de pouvoir traiter les données.

# Préparation et transformation des données

Une étape de préparation a été nécessaire afin de pouvoir répondre aux questions de l’exercice :

1. Ajout de la Colonne représentant la catégorie du vol (Militaire ou Commercial).
2. Ajout d’une colonne représentant le pays où l’écrasement est survenu.
3. Ajout d’une colonne représentant la surface de l’écrasement (Terre ou Mer).
4. Ajout de plusieurs colonnes représentant la destination du vol (ville, pays…)
5. Ajout d’une colonne représentant la raison du vol (entrainement, démonstration).
6. Ajout d’une colonne représentant la probabilité de survie.
7. Correction de plusieurs erreurs d’orthographe dans le nom des pays.
8. Remplacement des provinces canadiennes par le nom du pays
9. Remplacement des états américains par le nom du pays.
10. Regroupement par continent ou par hémisphère pour avoir des taxonomies niveau continent.

# Packages Utilisés

Les packages non inclus dans R base que nous avons utilisé pour ce devoir sont :

* ggmap
* plyr
* data.table

# Questions et Analyses

## Question 1 : « Nombre total d’écrasement à travers le temps »

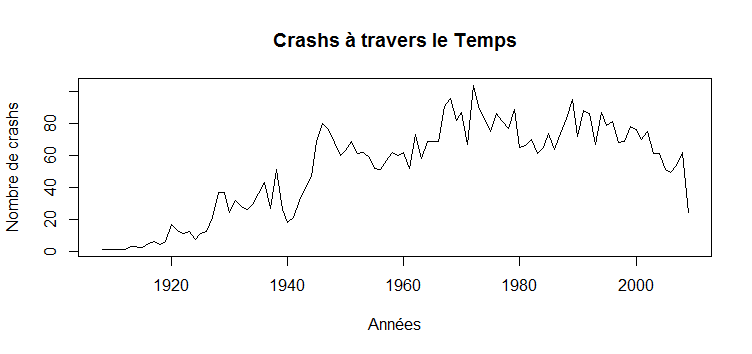
### Description

Cette analyse a pour but de déceler une tendance quant aux nombres d’écrasements qui ont eu lieu à travers le temps

### Approche de résolution

* Extraction de l’année (les 4 derniers caractères) de la variable « Date » à l’aide de la fonction substr et application de ce vecteur a la variable à « Annee »
* Combinaison du vecteur appelé « Annee » et d’une partie du jeu de données initial à « table\_temporaire » à l’aide de la fonction cbind.
* Appel de la fonction « trouver\_frequence » et application du résultat à la variable « x »
* Création d’un data.frame appelé « table\_occurence\_annee » avec « x » (vecteur de fréquence des dates) et le vecteur d’années sans duplicats
* Nomination des colonnes du data.frame
* Organisation de la « table\_occurence\_annee » en ordre croissant par date

### Résultats et interprétations



On peut voir qu’entre 1939 et 1945, le nombre d’écrasements d’avions a augmenté de façon significative. Des vols plus fréquents ainsi que des écrasements lors de la deuxième guerre mondial peuvent probablement expliqués cette augmentation. Par contre, on voit qu’à travers le temps, le nombre d’écrasements d’avions ne diminue pas. Cela semble plutôt stable, même si on pourrait être porté à croire que les avancées technologiques réduiraient le nombre d’écrasements.

### Améliorations possibles

Bien que le nombre de d’écrasements n’ait pas nécessairement baissé à travers le temps, il serait trompeur d’en tirer des conclusions. Plusieurs facteurs importants qui ne sont pas pris en compte dans notre fichier de données, tels que le nombre de vols durant une année, peuvent avoir affectés le nombre d’écrasements.

## Question 2 : « Nombre de fatalités par année à travers le temps»

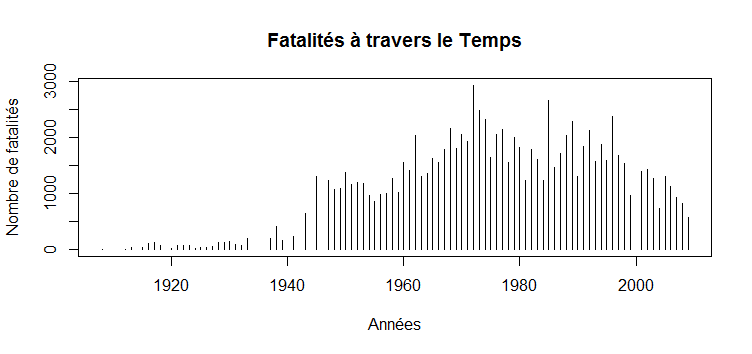
### Description

Dans cette analyse, on s’intéresse plutôt à connaître le nombre de fatalités qui ont eu lieu à travers le temps.

### Approche de résolution

* Création d’un data.frame avec quelques colonnes du jeu de données initiales et les années
* La définition du nom des colonnes du data.frame appelé table\_de\_fatalites\_par\_annee
* Utilisation d’une boucle pour trouver la somme du nombre de fatalités pour chaque année
* L’avant-dernière étape de préparation des données implique la création d’un data.frame avec chaque valeur de somme par années et les années
* La dernière étape de préparation des données comprend la création d’un graphique à l’aide la fonction « plot »

### Résultats et interprétations



Comme on peut voir dans le graphique, le nombre de fatalités semble suivre la même tendance que le graphique précédent « nombre d’écrasement à travers le temps ». La seule différence qui peut être notée c’est que le nombre de fatalités est plutôt faible entre les années 1940 et 1945. Cette période marqué par la guerre pourrait avoir été marqué par plusieurs écrasements d’avions de type militaire qui transportent souvent un ou deux passagers. Ceci pourrait potentiellement expliquer cette différence.

### Améliorations possibles

Les données auraient pu être divisées par type d’avions pour avoir une meilleure idée du nombre de fatalités pour les avions de type commerciaux. De plus, comme il a déjà été mentionné, avoir une idée du nombre de vols par année nous permettrait de faire un ratio. Et ceci nous permettrait de savoir s’il y a une amélioration technologique qui permet de réduire le nombre de fatalités lors des écrasements.

## Question 3 : « Nombre de fatalités par localisation»

### Description

Nous avons voulu savoir s’il y a certaines parties du monde qui sont plus touchés par des écrasements d’avions.

### Approche de résolution

* Utilisation d’une boucle pour trouver la somme du nombre de fatalités pour chaque localisation
* L’organisation de la table en ordre décroissant par nombre de fatalités par localisation
* L’extrait du top 50 des locations avec le plus grand nombre de fatalités par localisation
* L’utilisation d’une boucle pour transformer en « string » chacun des exemples de la variable « location » de la table top\_50\_location\_par\_fatalites
* La dernière étape de préparation des données comprend l’utilisation de la fonction « afficher\_map » pour faire afficher sur un map le top 50 des locations avec le plus grand nombre de fatalités.

### Résultats et interprétations

On peut voir sur le graphique le top 50 des endroits où il y a eu le plus grand nombres de fatalités. On peut voir qu’aux États-Unis, il y a une concentration importante de points rouge sur la côte est du pays. On peut facilement reconnaître le Brésil avec deux points rouges. Il fait partie des deux pays en Amérique latine à faire partie du top 50 des endroits avec le plus grand nombre de fatalités. Mis à part aux États-Unis, on voit que la plupart des fatalités les plus importantes ont pris place un peu partout dans le monde. Il est difficile de déceler une tendance précise dans ce cas-ci.

### 

### Améliorations possibles

Il aurait été intéressant d’utilisé un map qui affiche le nom des villes, pays et le nombre exact de fatalités. Malheureusement, le manque de connaissance technique de la librairie ggmap nous a permis de seulement offrir le map affiché ci-dessus. À noter que cinq valeurs qui commencent par un énoncé autre qu’un nom de ville n’affichent pas sur le map. Par exemple, les valeurs qui comment avec « Near, AtlanticOcean, off » n’ont pas retournés de coordonnées géographiques.

## Question 4 : « Nombre d’écrasements par localisation»

### Description

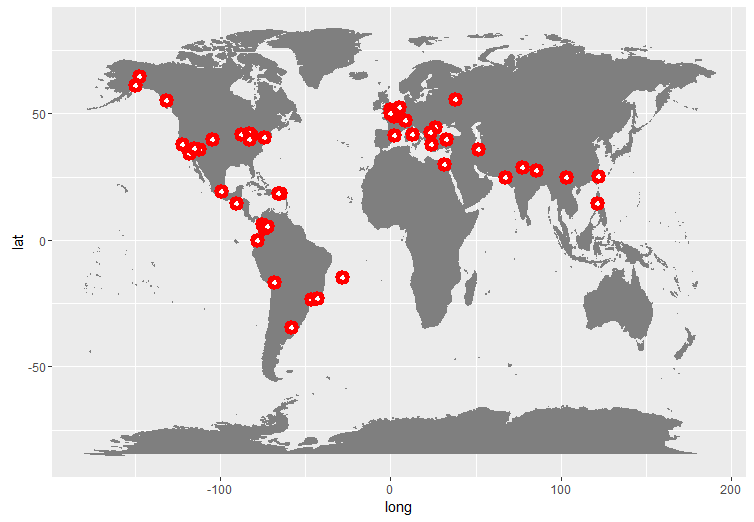
Afin de savoir s’il y a des zones plus dangereuses que d’autres dans le monde, on a voulu savoir s’il y avait plus d’écrasements dans des endroits spécifiques dans le monde. Cette question est une continuation de la question précédente.

### Approche de résolution

1. L’utilisation de la fonction trouver\_fréquence pour trouver le nombre de crashs par location
2. La création d’un data.frame pour combiner la fréquence par « location » et les «  locations »
3. L’utilisation de la fonction « order » pour mettre en ordre décroissant du nombre de crash par location
4. L’extrait du top 50 des « locations » avec la plus grande fréquence de nombres crashs
5. L’étape finale comprend l’utilisation de la fonction « afficher\_map » pour faire afficher sur un map les villes avec le plus grand nombre de crashs.

### Résultats et interprétation :

On peut voir que la plupart des écrasements d’avions ont eu lieu dans plusieurs villes d’Europe et des États-Unis. Ceci peut être expliqué par le fait que ces régions ont plus de vols qui traverse leur territoire. Il y a aussi le Brésil qui parait encore sur le map avec deux villes côte à côte. Ces deux villes sont probablement Rio de Janeiro et Sao Paulo qui sont les deux villes Brésiliennes les plus occupés en termes de trafic aérien.



### Amélioration possible :

À noter que la fonction afficher\_map omet d’afficher une valeur « AtlantiOcean, 110 miles West of Ireland ». Ceci est dû au fait que la valeur entrée n’est pas un nom de ville,

donc il est difficile de retrouver les coordonnées géographiques de cette valeur.

## Question 5 : « Quels sont les endroits les plus dangereux»

### Description

Le but de cette question était de regarder quels sont les endroits où des accidents d’avions ont été le plus recensés dans le passé.

### Approche de résolution

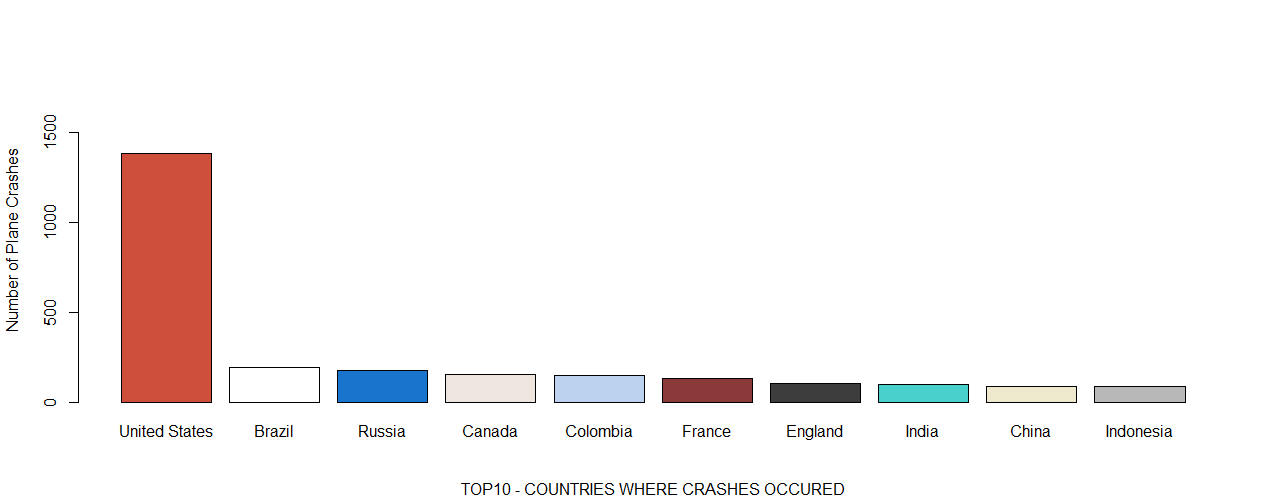
Nous nous basons principalement sur des statistiques de fréquences afin de terminer les destinations où les avions allaient, le pays où l’avion s’est écrasé selon la catégorie de l’avion.

1. Ajout de colonnes représentants les lieux des crashs et les destinations des avions lors des crashs. Ceci est fait dans la méthode apply. Ensuite la matrice est transposée et ensuite convertie en un dataframe.
2. Le pays est trouvé en « tokenizant » les données de la colonne Location. Le pays est habituellement le dernier élément du groupe de token. Dans le cas où le dernier élément est une province des États-Unis, une province du Canada ou un lieu dans la mer, la fonction states et les listes canada et sea\_words sont utilisés.
3. La fonction fix\_country\_names est utilisée pour corriger le nom des pays.
4. Après que les colonnes ont été ajoutées, répondre aux questions équivaut à faire des statistiques de fréquences à l’aide de la fonction count.

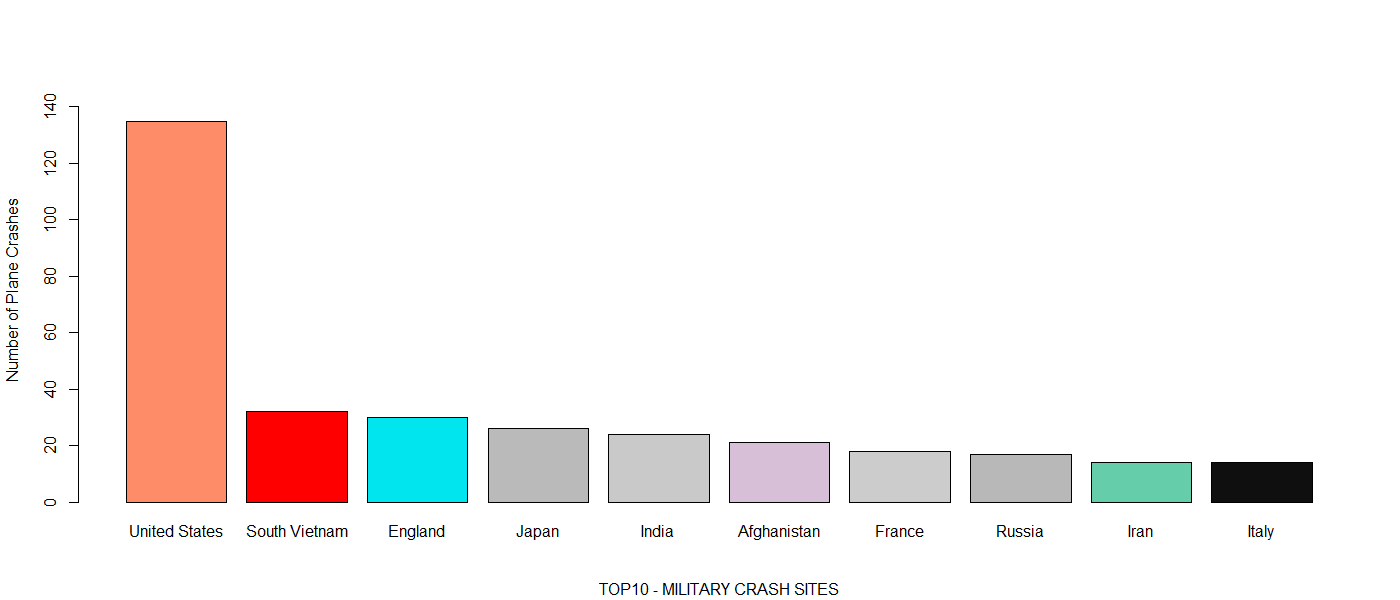
### Résultats et interprétations

Premièrement, la Figure 1 présente le top 10 des endroits où sont survenus des accidents d’avions. Nous voyons que les États-Unis sont en première place, suivit du Brésil, de la Russie et du Canada. Il est intéressant de remarquer que la plupart des pays dans ce TOP 10 sont des pays ayant un transport aérien développé (dans le sens de plusieurs décollages et arrivés par jour) et possédant dans la plupart des cas un constructeur d’avion.

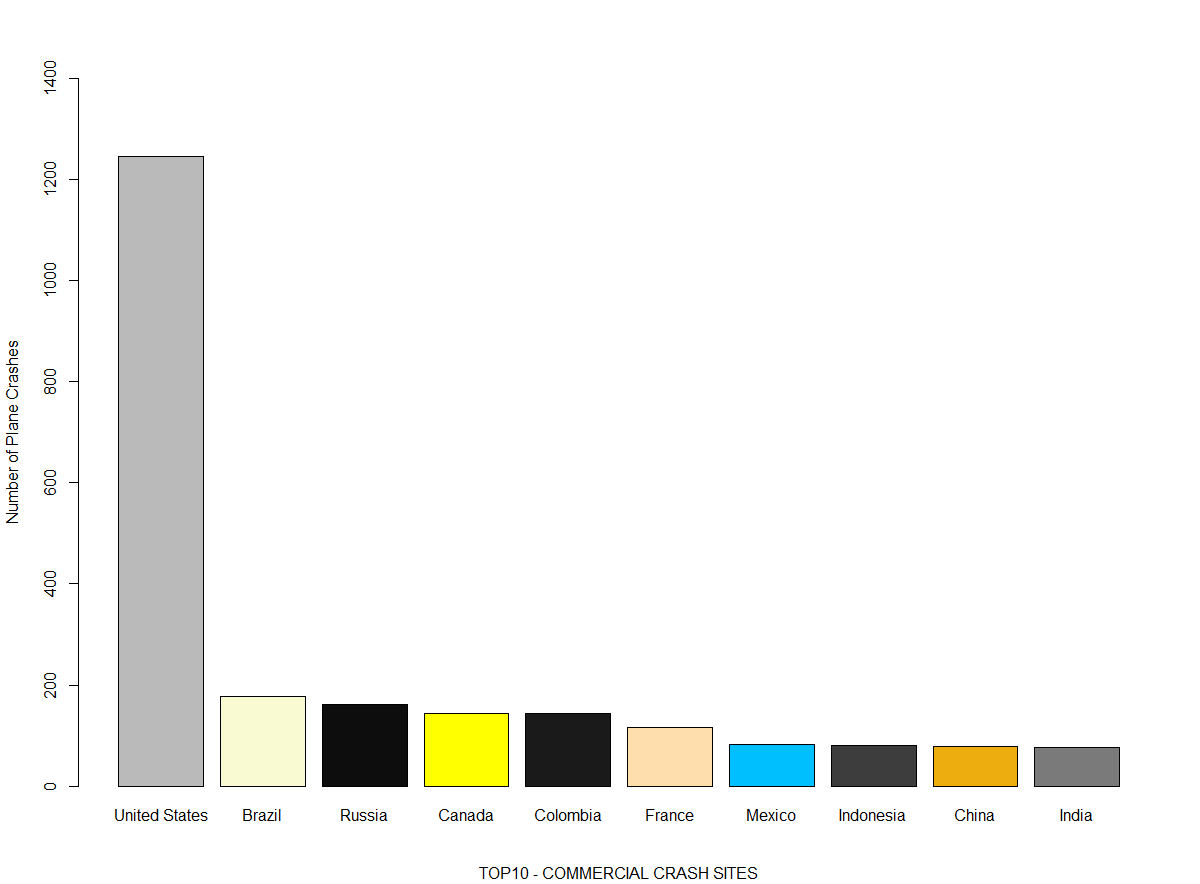
Un travail de suivi sera de vérifier s’il existe une corrélation entre le fait qu’un pays possède un constructeur aéronautique et le nombre de crash d’avion.



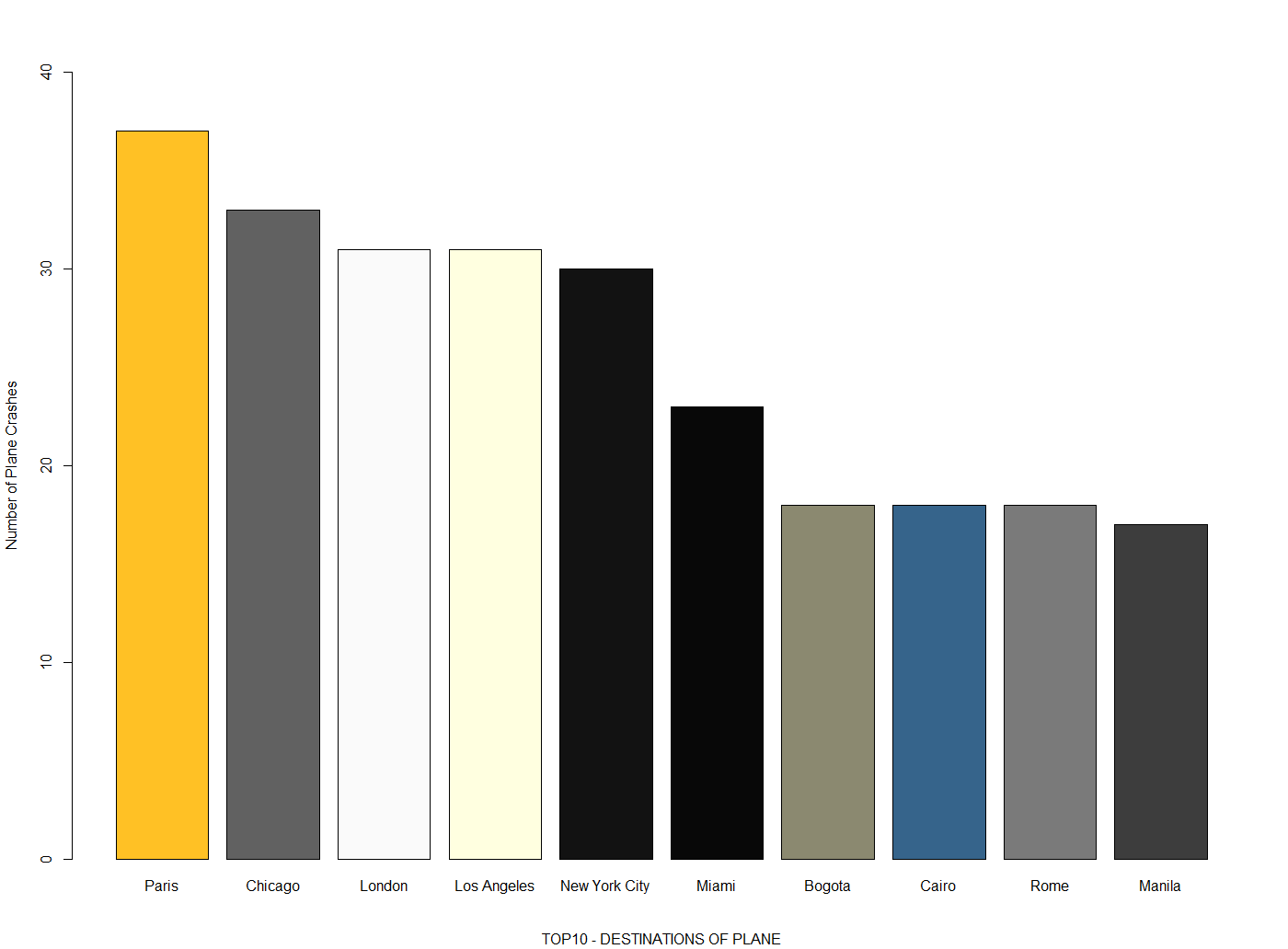
Dans les Figures suivantes, le nombre d’accidents est divisé selon que l’avion était un avion de type militaire ou commercial.

Ici, les États Unis sont toujours premiers, cependant, il est intéressant de voir que le deuxième pays pour les crashs militaires est le Sud Vietnam. Ceci nous démontre dans la collecte du jeu de données, plusieurs pays cachent l’information lorsque des crashs de nature militaires (ou même civil) surviennent. Il est important pour eux de toujours présenter leur pays sous un bon jour. D’après les auteurs, le Sud Vietnam occupe cette position du fait que les écrasements d’avions ont été répertoriés durant la guerre du Vietnam, ce qui n’est pas le cas pour plusieurs autres conflits.

Un des auteurs avait espéré que la Figure suivante démontrerait l’existence du fameux triangle des Bermudes <https://en.wikipedia.org/wiki/Bermuda_Triangle> . Ceci n’est pas le cas, les données compilées n’ont aucun pays de ce triangle parmi les endroits où surviennent le plus d’accidents d’avion.



La figure suivante nous montre que la plupart des accidents surviennent lors des vols reliant de grandes villes de ce monde. En effet, le sens commun nous dit que plus il y a de vols vers une destination, plus il y aura des accidents parmi les vols vers cette destination. Un exercice statistique serait de démontrer cette inférence

.

## Question 6 : « Quel est le risque selon les types d’avion »:

### Description

Le but de l’exercice est de déterminer le nombre d’écrasement par catégorie, soit commercial ou militaire.

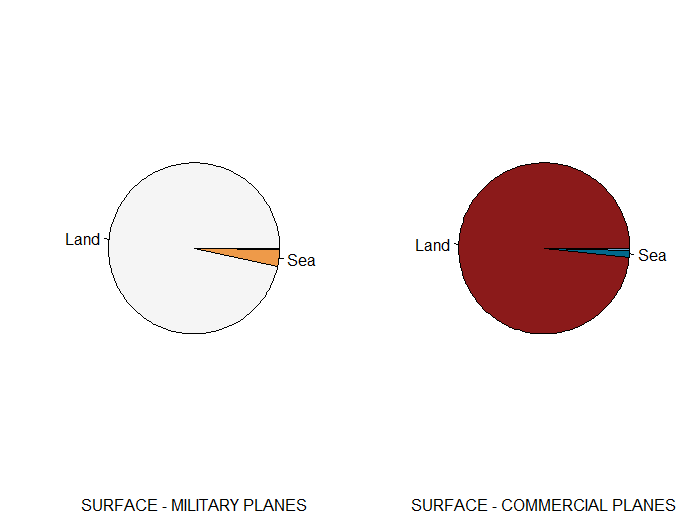
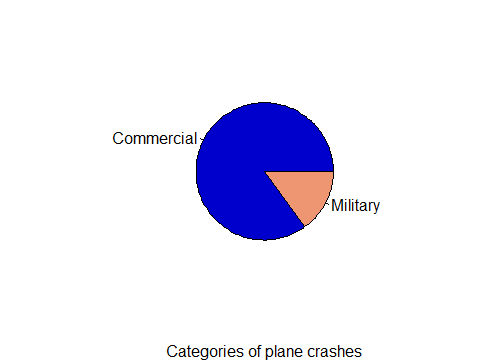
### Approche de résolution

Pour cette question des statistiques de fréquence sont utilisés. Nous avons préparé les données en ajoutant une colonne décrivant la catégorie d’avion.

1. Nous ajoutons une nouvelle colonne pour créer différencier les avions en catégories. Les avions sont les avions où le nom de l’opérateur contient des termes militaires, exemple : army, navy, air force, military…
2. Lorsque la colonne est ajoutée, répondre à la question équivaut à faire une statistique de fréquence avec la fonction count.

### Résultats et interprétations

Le jeu de données nous montre que la plupart des avions à être impliquer dans des écrasements sont des avions commerciaux. Il est important de mentionner qu’il existe un biais. En effet, comme mentionné plus haut dans le rapport, la plupart des pays ne signale pas lorsqu’un de leur avion militaire s’écrase. Il est aussi intéressant de remarquer que les accidents d’avions surviennent dans la grande majorité des cas sur la terre ferme.



## Question 7 : « Quels sont les pourcentages de survie selon le type d’avions?

### Description

Dans cette question, nous calculons quelle est la probabilité de survivre lors d’un accident d’avion.

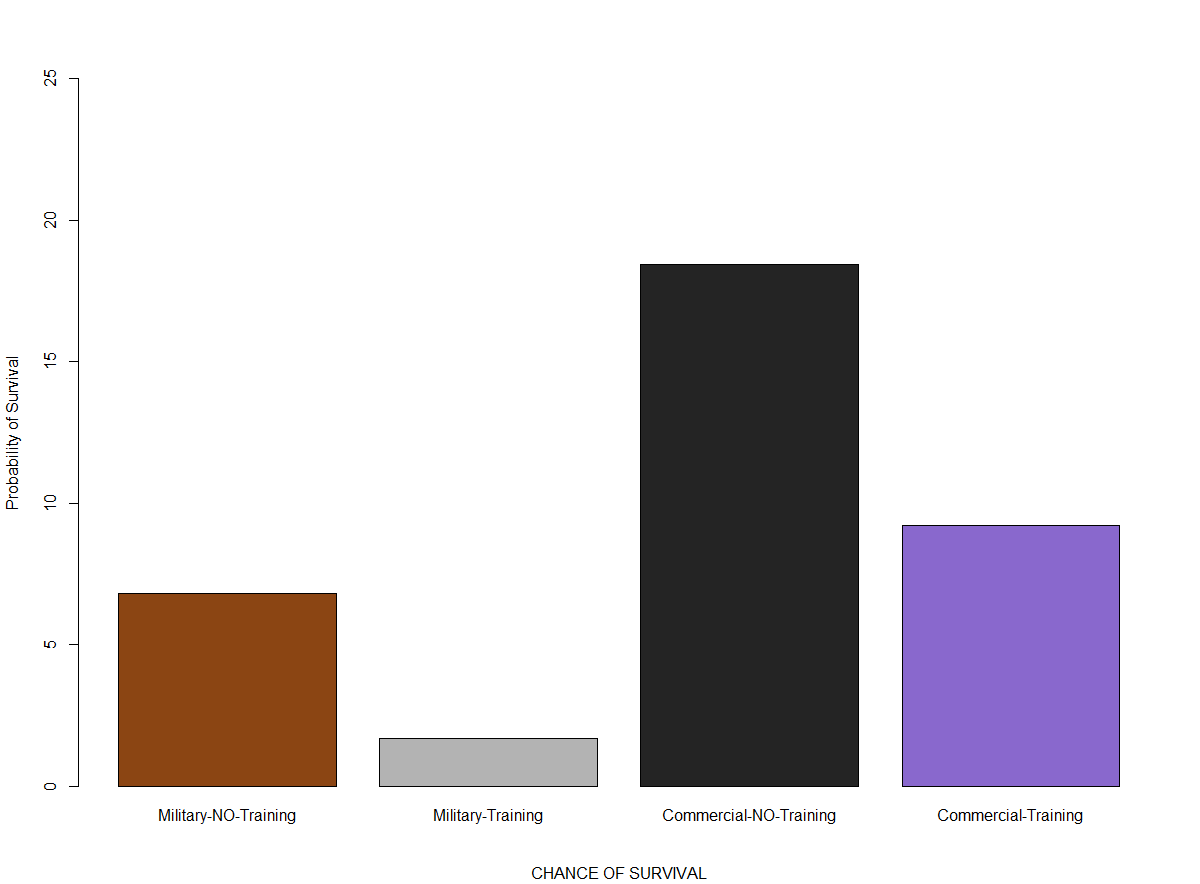
### Approche de résolution

Nous calculons le ratio de survie lors des accidents. Ce ratio est définie par :

1. Ici encore, nous ajoutons une colonne dans le code de la fonction apply où le pourcentage de survit est calculée pour chaque vol en utilisant la formule ci-dessus. Nous déterminons aussi si l’avion était en Training (entrainement) ou non.
2. Nous créons ensuite un dataframe correspond au type de l’avion et nous ajoutons si l’avion était en training ou pas.

### Résultats et interprétations

Le dernier graphique nous montre l’estimation de survie lors d’un crash selon que l’avion soit militaire ou commercial et que le crash survient lors d’un entrainement ou non.



## Question 8 : « Quelles sont les causes principales des écrasements d’avions»

### Description

Le but de cette question est d’avoir une idée globale des raisons pour lesquelles les Crashs arrivent.

### Approche de résolution

Pour pouvoir avoir une idée globale des raisons, voici l’approche adoptée :

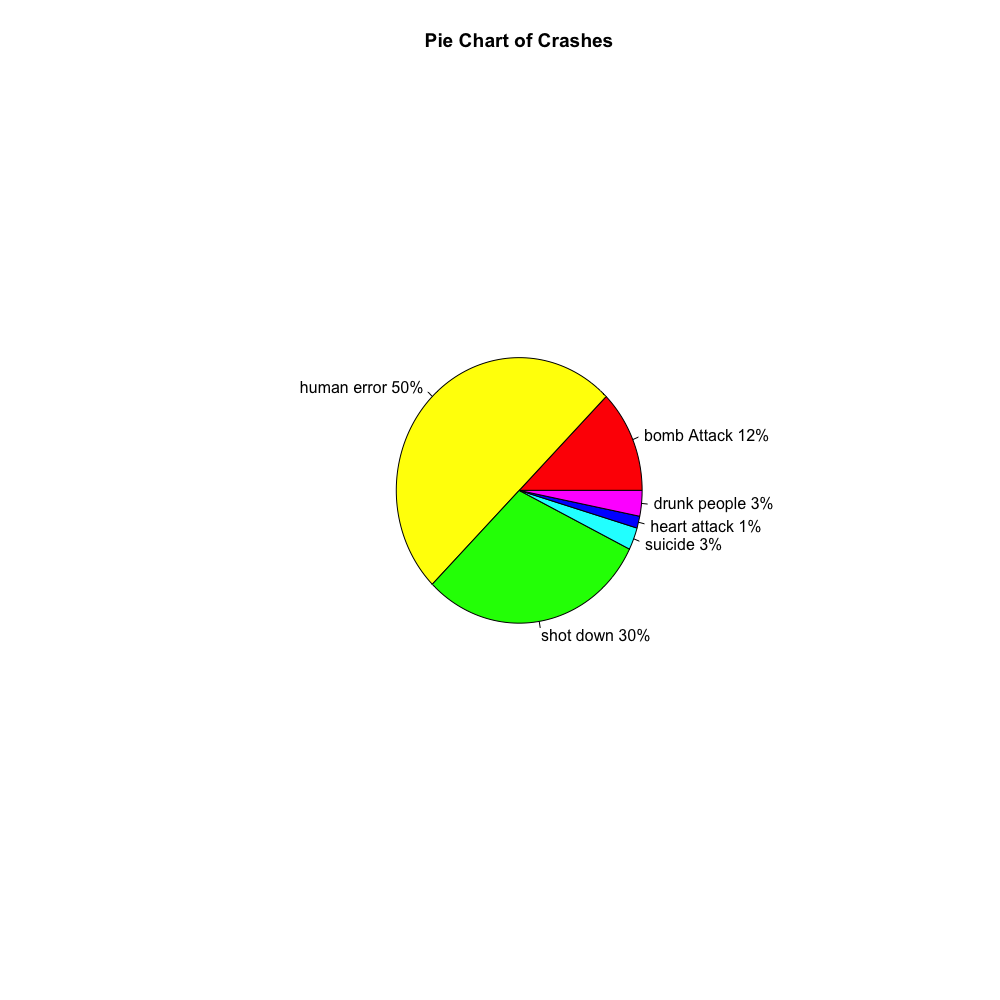
* Création des taxonomies pour les catégories :

BOMB\_ATTACK\_CATEGORIES, ERROR\_CATEGORIES, SHOT\_DOWN\_CATEGORIES, SHOT\_DOWN\_CATEGORIES, SHOT\_DOWN\_CATEGORIES DRUNK\_ATTACK\_CATEGORIES

* Parcours du data frame et Identification de catégorie pour chaque entrée
* L’approche choisie est une approche naïve qui consiste à trouver une série de mots clés dans la description du crash.

### Résultats et interprétations

Comme le résultat suivant le montre, la majorité (50%) des accidents présents dans le jeu de données sont causés par les erreurs humaines. Cette catégorie est suivie de près par la catégorie “Shot Down”. Ceci peux s’expliqué par la période des Crash de 1939-1945 (Guerre mondiale). La catégorie à plus faible proportion est celle des pilotes qui ont eu des maux de coeur pendant le vol. Ce qui explique pourquoi les tests de santé sont pris au sérieux par les compagnies aériennes.



### Améliorations possibles

L’approche de résolution choisie n’est pas optimale car elle peut être erronée. C’est une approche naïve. L’idéal serait d’aller vers une approche de “Text analysis”. Cette technique donnera beaucoup plus de résultats et serait beaucoup plus efficace. Une autre approche serait de réduire le jeu de données en enlevant tous les enregistrements pendant la période de guerre.

## Question 9 : « Quelle est la période de l’année la plus dangereuse et quelles sont les destinations les plus menaçantes ? »

### Description

Le but de cette question est de connaître la période de l’année la plus dangereuse ainsi que les destinations concernées.

### Approche de résolution

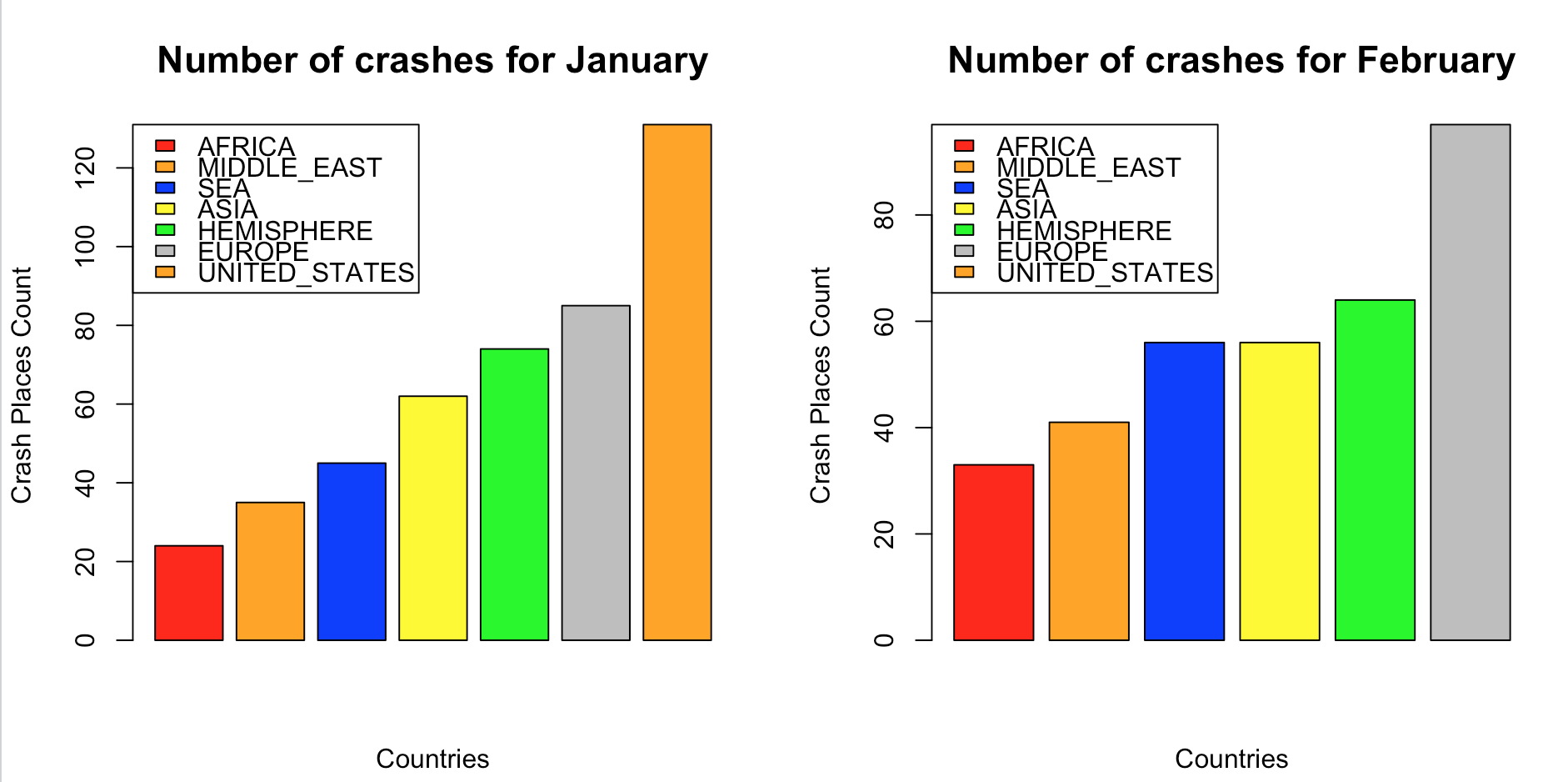
Elle se détaille comme suit :

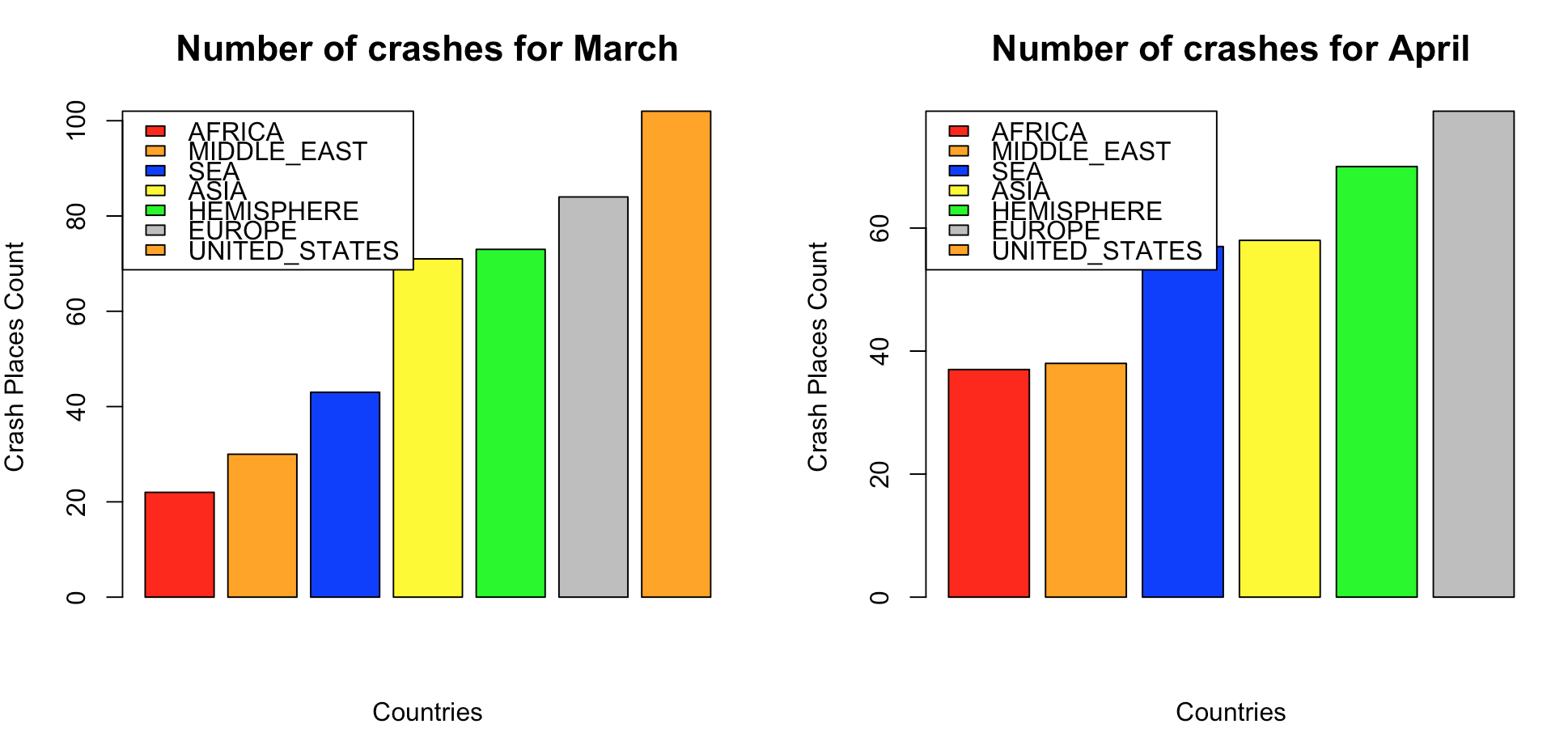
* Correction des noms de pays erronés
* Élimination des entrées vides ou manquantes
* Création des taxonomies pour les locations :

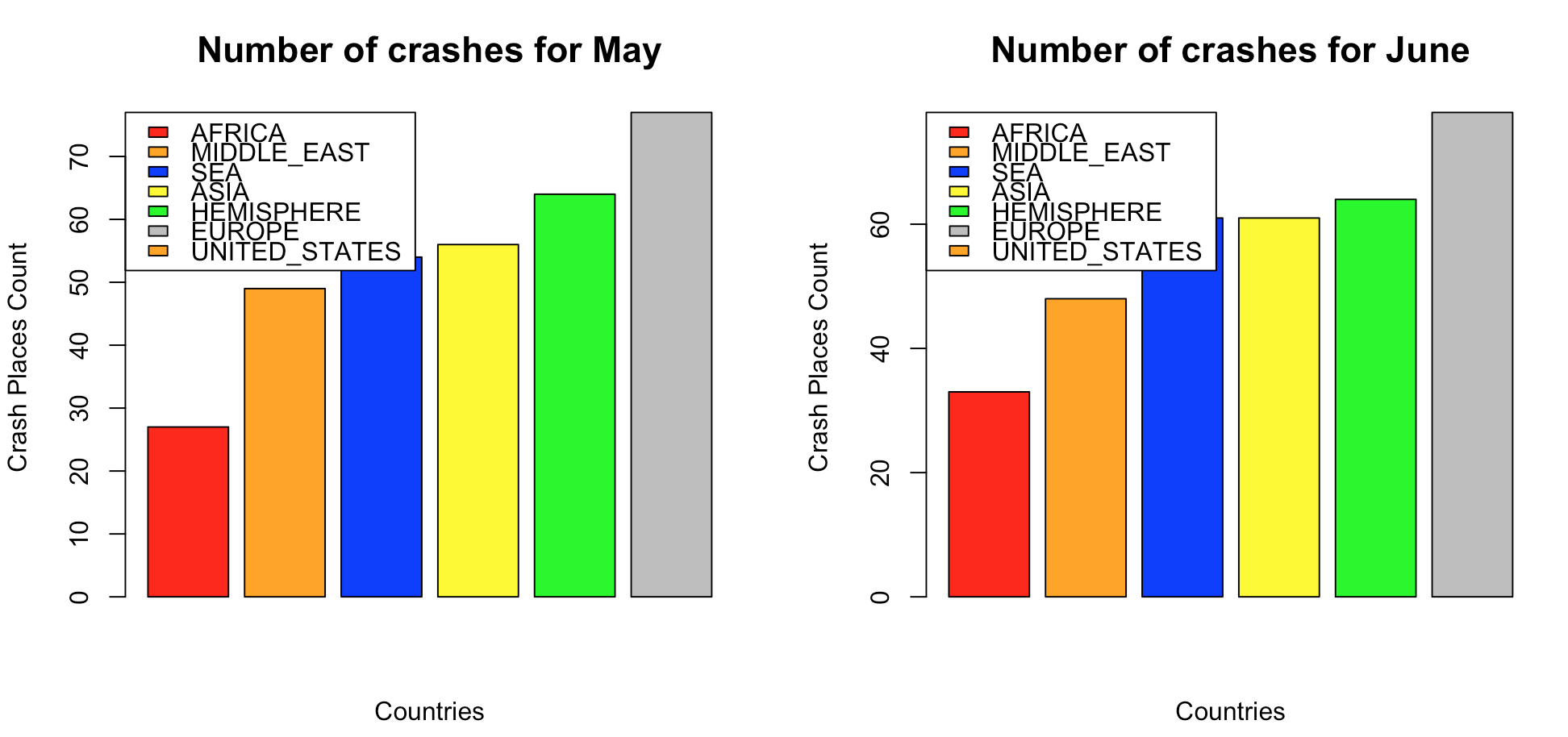
CANADA, AFRICAN\_COUNTRIES, WESTERN\_HEMISPHERE, SEA\_WORLD, EUROPE\_COUNTRIES EUROPE\_COUNTRIES, SOUTHEAST\_ASIA\_COUNTRIES, US\_CITIES, MIDDLE\_EAST

* Parcours du data frame et Identification de location pour chaque entrée
* L’approche choisie est une approche naïve qui consiste à trouver une série de mots clés dans la colonne location de l’entrée.

### Résultats et interprétation







Comme le montre les résultats, les records sont partagés entre les “UNITED\_STATES” et “EUROPE”. Ceci peut encore s’expliquer par la période des Crash dans le jeu de données (Guerre mondiale) ou encore par le fait que ces pays ont un plus de trafic aérien que les autres.

On peut voir que l’Afrique a eu très peu de crash. Ceci peut encore s’expliquer par le fait que le continent était très peu impliqué durant la deuxième Guerre Mondiale et par le fait que le trafic aérien a toujours été moins important sur ce continent.

### Améliorations possibles

Une amélioration consisterait à supprimer tous les enregistrements pendant la guerre mondiale, ou de se concentrer uniquement sur un type de vol, commercial ou militaire.

## Question 10 : « Quels sont les opérateurs qui ont fait le plus de victimes »

### Description

Le but de la question est de connaître les opérateurs qui ont fait plus de morts. Ceci pourrait influencer la compagnie aérienne avec laquelle on fait affaire lors de notre prochain vol.

### Approche de résolution

Elle se détaille comme suit :

* Élimination des entrées vides ou manquantes
* Création des taxonomies pour les opérateurs:

AIR\_FRANCE, AIR\_CANADA, TRANS\_CANADA, AIR\_ONTARIO, AIR\_INDIA, AIR\_CARAIBES, AIR\_MADAGASCAR, AIR\_NIAGARA, AIR\_GUADELOUPE, AIR\_AMERICA, AIR\_MALI,

DEUTSCHE\_LUFTHANSA, CHINA\_AIRLINES

* Parcours du data frame et Identification de l’opérateur pour chaque entrée
* L’approche choisie est une approche naïve qui consiste à trouver une série de mots clés dans la colonne Operateur de l’entrée.

### Résultats et interprétation

### 

Les résultats montrent que “AIR\_FRANCE” a eu plus de crash que les autres opérateurs. Ceci peut s’expliquer par le fait que “AIR\_FRANCE” a de plus gros avions que les autres opérateurs.L’observation qui surprend est celle de “AIR\_MADAGASGAR” car Madagascar fait partie de l’Afrique et l’Afrique n’avait pas de gros avions ni beaucoup de vols dans le passé.

### Améliorations possibles

* L’analyse est naïve dans la recherche de l’opérateur. Ceci peut être une source d’erreur dans l’analyse.
* Le type d’avion devrait aussi rentrer dans l’équation car plus un opérateur a de gros avions plus il a de chances d’avoir des bilans lourds quant au nombre de victimes.
* Investiguer pourquoi “AIR\_MADAGASCAR” a eu beaucoup de crash.

# Conclusion

En conclusion, bien que notre fichier de données réponde à tous nos critères de départ, nous pensons que nous aurions dû être plus sévères en ce qui concerne la qualité du fichier de données. En effet, aucune description des variables n’a été fournie avec le fichier de données. Plusieurs données étaient manquantes. De plus, la source de la base de données nous est inconnue, ce qui nous pousse à nous questionner sur la qualité et la véracité des données. Pour de meilleurs statistiques sur le sujet, le Bureau des Archives d’Accident d’Avions tient une base de données en ligne qui permet d’explorer le sujet.

Source :

Base de données du Bureau des Archives sur Accidents d’Avions :

<http://www.baaa-acro.com/general-statistics/>

Le lien de notre fichier de données :

https://www.kaggle.com/saurograndi/airplane-crashes-since-1908