

**MỤC LỤC**

[1 giới thiệu dự án](#_gjdgxs) 5

[1.1 giới thiệu](#_30j0zll) 5

[1.2 yêu cầu của công ty](#_1fob9te) 5

[1.3 lập kế hoạch dự án](#_3znysh7) 10

[2 phân tích yêu cầu tuyển dụng](#_tyjcwt) 13

[2.1 phân tích yêu cầu](#_3dy6vkm) 13

[2.2 câu chuyện dữ liệu](#_26in1rg) 14

[2.2.1 đặt vấn đề](#_lnxbz9) 14

[2.2.2 xác định câu chuyện](#_35nkun2) 15

[2.2.3 xác định rõ đối tượng](#_1ksv4uv) 15

[2.2.4 xác định câu chuyện chi tiết](#_44sinio) 17

[2.2.5 trình bày dữ liệu](#_2jxsxqh) 17

[2.2.6 những điều cần lưu ý](#_3j2qqm3) 19

[2.3 kiến trúc hệ thống](#_1y810tw) 19

[2.3.1 kiến trúc](#_4i7ojhp) 19

[2.3.2 giải thích](#_2xcytpi) 20

[2.4 giải thích về bộ dữ liệu khách hàng](#_1ci93xb) 22

[2.4.1 các khái niệm](#_3whwml4) 22

[2.4.2 các trường dữ liệu](#_2bn6wsx) 22

[3 làm sạch và chuyển đổi dữ liệu](#_41mghml) 22

[3.1 chuẩn bị dữ liệu](#_2grqrue) 22

[3.1.1 giải pháp lưu trữ dữ liệu](#_vx1227) 22

[3.1.2 giải pháp phân bố dữ liệu](#_3fwokq0) 25

[3.2 làm sạch dữ liệu](#_nmf14n) 29

[3.2.1 các vấn đề ảnh hưởng tới dữ liệu](#_37m2jsg) 29

[3.2.2 các tiêu chí đánh giá chất lượng dữ liệu](#_1mrcu09) 30

[3.2.3 các bước làm sạch dữ liệu](#_111kx3o) 32

[3.3 chuyển đổi dữ liệu](#_4du1wux) 34

[4 xử lý dữ liệu](#_meukdy) 35

[4.1 chuẩn hóa dữ liệu](#_36ei31r) 35

[4.1.1 trình bày các bước chuẩn hóa trong dự án](#_1ljsd9k) 35

[4.2 mô hình hóa dữ liệu](#_45jfvxd) 35

[4.2.1 các loại mô hình hóa](#_2koq656) 35

[4.2.2 các tiêu chí đánh giá mô hình dữ liệu](#_zu0gcz) 37

[4.2.3 trình bày các bước mô hình hóa](#_3jtnz0s) 41

[4.3 xử lý dữ liệu dax](#_j8sehv) 40

[4.3.1 measure](#_338fx5o) 41

[4.3.2 calculated column](#_1a346fx) 43

[4.3.3 filter](#_3w19e94) 45

[5 trực quan hóa dữ liệu](#_1pgrrkc) 45

[5.1 các kỹ thuật trực quan hóa](#_49gfa85) 46

[5.2 các nguyên tắc trực quan hóa](#_2olpkfy) 48

[5.4 trình bày tạo các report cho dự án](#_1h65qms) 59

[5.4.1 tạo visual thống kê chi tiết](#_415t9al) 59

[6 xây dựng báo cáo](#_1au1eum) 59

[6.1 dashboard và report](#_3utoxif) 59

[6.2 xây dựng báo cáo](#_29yz7q8) 60

[6.2.1 dashboard và report](#_p49hy1) 60

[6.2.2 dashboard](#_393x0lu) 60

[6.2.3 bookmark](#_2ne53p9) **66**

7 [kết luận](#_26sx1u5) 67

7[.1 báo cáo](#_ly7c1y) 67

7[.1.1 các bước viết báo cáo](#_35xuupr) 67

7[.1.2 tổng](#_1l354xk) hợp 68

7[.2 khó khăn](#_452snld) 69

7[.3 thuận lợi](#_2k82xt6) 69

7[.4 hướng phát triển](#_zdd80z) 70

8 [tổng kết](#_3jd0qos) 73

**MỤC LỤC BẢNG**

[**1 giới thiệu dự án 8**](#_gjdgxs)

[1.1 giới thiệu 8](#_30j0zll)

[1.2 yêu cầu của công ty 8](#_1fob9te)

[**- Mục tiêu chính của việc thu thập dữ liệu là để phân tích tuyển dụng trong ngành công nghệ thông tin (CNTT) tại Việt Nam. 11**](#_wvnt4e5v92wx)

[1.3 lập kế hoạch dự án 13](#_3znysh7)

[**2 phân tích yêu cầu tuyển dụng 15**](#_tyjcwt)

[2.1 phân tích yêu cầu 15](#_3dy6vkm)

[2.2 câu chuyện dữ liệu 16](#_26in1rg)

[2.2.1 đặt vấn đề 16](#_lnxbz9)

[Khởi đầu và vấn đề của Hùng 16](#_7pq0vzek214q)

[Phát hiện bản tin phân tích xu hướng tuyển dụng IT 16](#_ta2xc46zk1w2)

[Phân bố việc làm theo địa lý 17](#_cfe168a4tp9y)

[Mâu thuẫn trong mức lương và cơ hội ngành nghề 17](#_92lxg93mwxdy)

[Niềm hy vọng và kế hoạch của Hùng 17](#_qizwh883jrm1)

[2.2.2 xác định câu chuyện 17](#_35nkun2)

[2.2.3 xác định rõ đối tượng 18](#_1ksv4uv)

[2.2.4 xác định câu chuyện chi tiết 19](#_44sinio)

[2.2.5 trình bày dữ liệu 20](#_2jxsxqh)

[2.2.6 những điều cần lưu ý 21](#_3j2qqm3)

[2.3 kiến trúc hệ thống 21](#_1y810tw)

[2.3.1 kiến trúc 21](#_4i7ojhp)

[2.3.2 giải thích 22](#_2xcytpi)

[2.4 giải thích về bộ dữ liệu tuyển dụng 23](#_1ci93xb)

[2.4.1 các khái niệm 23](#_3whwml4)

[2.4.2 các trường dữ liệu 23](#_2bn6wsx)

[**3 làm sạch và chuyển đổi dữ liệu 24**](#_41mghml)

[3.1 chuẩn bị dữ liệu 24](#_2grqrue)

[3.1.1 giải pháp lưu trữ dữ liệu 24](#_vx1227)

[3.1.2 giải pháp phân bố dữ liệu 27](#_3fwokq0)

[3.1.2.1 Ý nghĩa việc phân bố dữ liệu 28](#)

[3.1.2.2 Trình bày cách phân bố dữ liệu 29](#)

[3.2 làm sạch dữ liệu 31](#_nmf14n)

[3.2.1 các vấn đề ảnh hưởng tới dữ liệu 31](#_37m2jsg)

[3.2.2 các tiêu chí đánh giá chất lượng dữ liệu 32](#_1mrcu09)

[3.2.2.1 Các tiêu chí 32](#_46r0co2)

[3.2.2.2 Tiêu chí áp dụng trong dự án 32](#_2lwamvv)

[3.2.3 các bước làm sạch dữ liệu 34](#_111kx3o)

[3.2.3.1 Trình bày các bước làm sạch 34](#)

[3.2.3.2 Trình bày các bước làm sạch trong phạm vi dự án 34](#)

[3.3 chuyển đổi dữ liệu 37](#_4du1wux)

[các kỹ thuật chuyển đổi 37](#_184mhaj)

[**4 xử lý dữ liệu 38**](#_meukdy)

[4.1 chuẩn hóa dữ liệu 38](#_36ei31r)

[4.1.1 trình bày các bước chuẩn hóa trong dự án 38](#_1ljsd9k)

[4.2 mô hình hóa dữ liệu 38](#_45jfvxd)

[4.2.1 các loại mô hình hóa 38](#_2koq656)

[4.2.2 các tiêu chí đánh giá mô hình dữ liệu 40](#_zu0gcz)

[4.2.3 trình bày các bước mô hình hóa 41](#_3jtnz0s)

[4.3 xử lý dữ liệu dax 44](#_j8sehv)

[4.3.1 measure 44](#_338fx5o)

[4.3.1.1 Tạo bang data 44](#)

[4.3.2 calculated column 46](#_1a346fx)

[4.3.2.1 Bảng Data (Fact) 46](#_3u2rp3q)

[4.3.2.2 Bảng bằng cấp 46](#_odc9jc)

[4.3.2.3 Bảng công ty 47](#_1nia2ey)

[4.3.2.4 Bảng công việc 47](#_4kx3h1s)

[4.3.2.5 Bảng phúc lợi 48](#_2rrrqc1)

[4.3.2.6 Bảng nhóm công việc 48](#_3im3ia3)

[4.3.3 filter 49](#_3w19e94)

[4.3.3.1 Tạo filter chọn nhóm công việc, thành phố, cấp độ phúc lợi, ngành nghề 49](#)

[4.3.3.2 Tạo filter chọn Ngày 49](#_qbtyoq)

[**5 trực quan hóa dữ liệu 49**](#_1pgrrkc)

[5.1 các kỹ thuật trực quan hóa 49](#_49gfa85)

[5.2 các nguyên tắc trực quan hóa 51](#_2olpkfy)

[5.3 trình bày tạo các report cho dự án 52](#_1h65qms)

[5.3.1 tạo visual thống kê chi tiết 52](#_415t9al)

[5.3.1.1 Trang tổng quát 52](#)

[5.3.1.2 Trang lương 53](#)

[5.3.1.3 Trang kinh nghiệm và cấp bậc 56](#)

[5.3.1.4 Trang ngành nghề và nhóm công việc 59](#)

[5.3.1.5 Trang thành phố và ngoại ngữ 60](#_sabnu4)

[**6 xây dựng báo cáo 62**](#_1au1eum)

[6.1 dashboard và report 62](#_3utoxif)

[6.2 xây dựng báo cáo 63](#_29yz7q8)

[6.2.1 dashboard và report 63](#_p49hy1)

[6.2.2 dashboard 63](#_393x0lu)

[Dashboard tổng quan (overview dashboard) 64](#)

[Dashboard phân tích mức lương (Salary) 65](#)

[Dashboard phân tích kinh nghiệm và cấp bậc 66](#)

[Dashboard phân tích ngành nghề và nhóm công việc 67](#)

[-Dashboard phân tích thành phố và ngoại ngữ 68](#_uacuyob3qz0o)

[6.2.3 bookmark 69](#_2ne53p9)

[6.2.3.1 Tạo bookmark overview 69](#)

[6.2.3.2 Tạo bookmark chuyển tiếp 69](#)

[**7 kết luận 70**](#_26sx1u5)

[7.1 báo cáo 70](#_ly7c1y)

[7.1.1 các bước viết báo cáo 70](#_35xuupr)

[7.1.2 tổng hợp 71](#_1l354xk)

[7.2 khó khăn 72](#_452snld)

[7.3 thuận lợi 72](#_2k82xt6)

[7.4 hướng phát triển 73](#_zdd80z)

[Mở rộng phạm vi phân tích: 73](#_3yuz3x6zgsh)

[Nâng cao chất lượng dữ liệu: 73](#_z3eu3q265552)

[Phát triển mô hình dự đoán: 73](#_jghizi6m05ip)

[Cải thiện trực quan hóa dữ liệu: 73](#_p7dl1svk70tl)

[Nâng cao trải nghiệm người tìm việc: 74](#_75qi59m3dktm)

[Tối ưu hóa hoạt động tuyển dụng: 74](#_g7pm1mu9ro7f)

[Mở rộng thị trường và danh mục ngành nghề: 74](#_klmqjddim9xw)

[Nâng cao hiệu quả hoạt động tuyển dụng: 74](#_v45yebhdzsmn)

[**8 tổng kết 76**](#_3jd0qos)

**MỤC LỤC HÌNH ẢNH**

[Hình](#_1v1yuxt) 2.1: Mô hình kiến trúc hệ thống18

[Hình 4.1: Mối quan hệ](#_1yyy98l) 39

[Hình 4.2: Danh sách các measure được tạo trong báo cáo](#_2y3w247) 41

[Hình 4.3:](#_1d96cc0) Measure tính kinh nghiệm thực tập sinh 41

[Hình 4.4: Measure tính lương trung bình](#_3x8tuzt) 41

[Hình 4.5: Công thức tính lương chênh lệch](#_2ce457m) 41

[Hình 4.6: Công thức tính lương trung bình theo thành phố](#_rjefff)  42

[Hình 4.7: Measure tính lương trung bình theo cấp bậc](#_3bj1y38)  42

[Hình 4.8: Measure top 3 thành phố với mức lương cao nhất](#_1qoc8b1) 42

[Hình 4.9: Measure tính số lượng lương theo cấp bậc](#_4anzqyu) 42

[Hình 4.10: Bảng Data](#_2pta16n) 43

[Hình 4.11: Bảng bằng cấp](#_14ykbeg) 43

[Hình 4.12: Bảng công ty](#_3oy7u29) 44

[Hình 4.13: Bảng](#_243i4a2) công việc 44

[Hình 4.14: Bảng phúc lợi](#_1idq7dh) 45

[Hình 4.15: Bảng nhóm công việc](#_42ddq1a) 45

[Hình 4.16:](#_2hio093) Bộ lọc chọn nhóm công việc, thành phố, cấp độ phúc lợi, ngành nghề46

[Hình 4.17:Bộ lọc chọn ngày](#_wnyagw) 46

[Hình 5.1:](#_3nqndbk) Card tổng số công việc, trung bình lương min,trung bình lương max,trung bình kinh nghiệm 49

[Hình 5.2: Biểu đồ chênh lệch lương max và min](#_22vxnjd) 50

[Hình 5.3: Biểu đồ số lượng công việc theo tháng](#_i17xr6) 50

[Hình 5.4:](#_320vgez) Card thống kê về số bản tin có lương cứng , lương cạnh tranh , trung bình lương min ,trung bình lương max 51

[Hình 5.5: Biểu đồ thống kê lương max và lương min theo nhóm công việ*c*](#_2gb3jie) 52

[Hình 5.6: Biểu đồ so sánh tỷ lệ lương cạnh tranh và lương trung bình](#_vgdtq7) 52

[Hình 5.7: Biểu đồ lương theo số năm kinh nghiệm](#_3fg1ce0) 53

[Hình 5.8: Biểu đồ lương trung bình theo top 10 thành phố](#_1ulbmlt) 53

[Hình 5.9: Card thống kê về kinh nghiệm của thực tập sinh](#_4ekz59m) 54

[Hình 5.10: Biểu đồ phân tích số lượng công việc theo bậc](#_2tq9fhf) 54

[Hình 5.11: Biểu đồ phân tích kinh nghiệm trung bình theo từng cấp bậc](#_18vjpp8) 55

[Hình 5.12: Biểu đồ phân tích top thành phố yêu cầu kinh nghiệm trung bình thấp nhất](#_3sv78d1) 55

[Hình 5.13: Biểu đồ thể hiện top ngành nghề có lương cao nhất](#_280hiku) 56

[Hình 5.14: Biểu đồ thể hiện top nhóm công việc có lương cao nhất](#_n5rssn) 56

[Hình 5.15: Biểu đồ top 10 ngành nghề theo số lượng công việc](#_375fbgg) 57

[Hình 5.16: Biểu đồ thể hiện nhóm công việc theo số lượng công việc](#_1maplo9) 57

[Hình 5.17:](#_46ad4c2) visual thể hiện tỉ lệ việc làm theo ngoại ngữ58

[Hình 5.18:](#_2lfnejv) Visual thể hiện mức lương trung bình theo ngoại ngữ58

[Hình 5.19:](#_10kxoro) Biểu đồ thể hiện top 5 thành phố theo số lượng công việc59

Hình 5.20: Biểu đồ thể hiện top thành phố có lương cao nhất 59

[Hình 6.1:](#_488uthg) Overview 61

Hình 6.2: Phân tích mức lương 62

Hình 6.3: Phân tích kinh nghiệm và cấp bậc 63

Hình 6.4: Phân tích ngành nghề và nhóm công việc 64

Hình 6.5: Phân tích thành phố và ngoại ngữ 65

Hình 6.6: Bookmark overview 66

Hình 6.7: Bookmark chuyển tiếp 66

# giới thiệu dự án

## giới thiệu

Dự án Phân tích Tuyển dụng Ngành Công nghệ Thông tin (CNTT) tại Việt Nam được thực hiện nhằm mang đến cái nhìn toàn diện về thị trường lao động trong lĩnh vực công nghệ, trong bối cảnh chuyển đổi số diễn ra mạnh mẽ và nhanh chóng. Ngành CNTT hiện đang là một trong những lĩnh vực có nhu cầu nhân lực cao nhất, với các vị trí như lập trình viên, chuyên gia AI/ML, chuyên gia bảo mật, chuyên viên phân tích dữ liệu và quản lý dự án ngày càng được săn đón. Sự cạnh tranh trong việc thu hút nhân tài không chỉ diễn ra giữa các doanh nghiệp trong nước mà còn mở rộng ra quốc tế, khi các công ty nước ngoài tìm kiếm nhân lực chất lượng từ Việt Nam.

Dự án sử dụng bộ dữ liệu phong phú, được thu thập từ các nền tảng tuyển dụng hàng đầu như TopCV, Vietnamworks, vieclam24h và LinkedIn trong giai đoạn từ năm 2024. Dữ liệu bao gồm danh sách vị trí tuyển dụng, yêu cầu kỹ năng, mức lương, kinh nghiệm cần thiết và sự phân bổ địa lý của các cơ hội việc làm, cùng với thông tin về hồ sơ ứng viên. Với sự hỗ trợ của các công cụ phân tích dữ liệu hiện đại, dự án không chỉ đánh giá hiện trạng tuyển dụng mà còn làm rõ các xu hướng kỹ năng, mức độ cạnh tranh trong ngành, và sự khác biệt về yêu cầu công việc giữa các khu vực địa lý như Hà Nội, TP.HCM và các tỉnh thành khác.

Mục tiêu của dự án bao gồm việc hỗ trợ doanh nghiệp xác định chiến lược tuyển dụng hiệu quả, tối ưu hóa nguồn lực và giảm thiểu chi phí tuyển dụng. Đồng thời, dự án cũng cung cấp thông tin giá trị cho người lao động, giúp họ hiểu rõ hơn về nhu cầu thị trường, từ đó định hướng phát triển kỹ năng phù hợp và cải thiện cơ hội việc làm. Ngoài ra, dự án còn hướng tới việc phân tích các yếu tố thu hút ứng viên, đánh giá mức lương trung bình, phúc lợi và các chương trình đào tạo phù hợp với xu hướng công nghệ mới như AI, blockchain, điện toán đám mây, và big data.

Dự án không chỉ phục vụ mục tiêu ngắn hạn của các doanh nghiệp và ứng viên mà còn góp phần xây dựng một hệ sinh thái tuyển dụng bền vững, thúc đẩy sự phát triển lâu dài của ngành CNTT Việt Nam. Thông qua việc cung cấp các phân tích chuyên sâu và đề xuất thực tiễn, dự án hy vọng sẽ góp phần nâng cao năng lực cạnh tranh của ngành CNTT trên thị trường quốc tế, đồng thời giúp Việt Nam khẳng định vị thế là trung tâm công nghệ hàng đầu trong khu vực.

## yêu cầu của công ty

**Về mặt dữ liệu**

Về mặt dữ liệu, hệ thống tuyển dụng cung cấp thông tin chi tiết về các vị trí tuyển dụng, doanh nghiệp, yêu cầu và quyền lợi dành cho ứng viên. Bộ dữ liệu hiện tại bao gồm nhiều bảng liên kết, mỗi bảng mang thông tin quan trọng để hỗ trợ phân tích và tối ưu hóa chiến lược tuyển dụng. Cụ thể:

| **Cấp bậc** | | Mã bảng cấp | Bằng cấp | | --- | --- |   dùng để phân loại các cấp bậc trong hệ thống. Những thông tin này có thể hỗ trợ trong việc đánh giá mức độ thăng tiến và các phúc lợi tương ứng với mỗi cấp bậc |
| --- | --- | --- | --- |
| **Data** | | Cấp bậc | Cấp độ phúc lợi | Chênh Lệch | Kinh nghiệm | | --- | --- | --- | --- | | Lương cạnh tranh | Lương Max | Lương Min | Lương TB |   Những dữ liệu này cung cấp cái nhìn tổng quan về các điều kiện và yêu cầu của công việc, giúp doanh nghiệp so sánh và phân tích các mức lương, phúc lợi và kinh nghiệm cần thiết cho các vị trí công việc |
| **Phúc lợi** | | Mã phúc lợi | Cấp độ phúc lợi | | --- | --- |   thông tin này giúp xác định các loại phúc lợi mà nhân viên có thể nhận được, dựa trên cấp độ công việc hoặc thâm niên |
| **Nhóm công việc** | | Mã nhóm công việc | Nhóm công việc | | --- | --- |   sử dụng để phân loại và nhóm các công việc theo từng nhóm lĩnh vực hoặc loại hình công việc cụ thể. |
| **Công việc** | | Mã công việc | Tên công việc | Ngày cập nhật | Kinh nghiệm | | --- | --- | --- | --- | | Lương cạnh tranh | Lương max | Lương min |  |   Dữ liệu trong bảng này giúp quản lý và phân tích các yêu cầu, mức lương và các tiêu chí khác đối với từng công việc cụ thể |
| **Công ty** | | Mã công ty | Tên công ty | | --- | --- |   Thông tin này là cơ sở để phân loại và quản lý các công ty trong hệ thống, giúp phân tích mối quan hệ giữa các công ty và các công việc trong các lĩnh vực khác nhau. |

*Bảng 1.1: Bảng thống kê về dữ liệu*

Bộ dữ liệu này bao gồm hàng trăm nghìn bản ghi và đòi hỏi phải đảm bảo chất lượng để có thể khai thác tối đa. Dữ liệu cần phải được làm sạch, chuẩn hóa và mô hình hóa trước khi phân tích. Quá trình làm sạch sẽ bao gồm loại bỏ các bản ghi trùng lặp, xử lý các giá trị thiếu và xác định các ngoại lệ để đảm bảo độ chính xác. Việc chuẩn hóa dữ liệu giúp duy trì sự nhất quán trong toàn bộ hệ thống, từ định dạng thời gian đến chuẩn hóa tên cột. Cuối cùng, dữ liệu sẽ được mô hình hóa để phù hợp với các phân tích cụ thể, giúp tối ưu hóa hiệu suất truy vấn và giảm thiểu thời gian xử lý.

**Quản lý và lưu trữ**

* Cần xây dựng hệ thống quản lý và lưu trữ dữ liệu hiệu quả để dễ dàng truy cập và sử dụng khi cần thiết.
* Xây dựng hệ thống quản lý dữ liệu: Hệ thống này phải dễ dàng sử dụng và truy cập để đảm bảo rằng dữ liệu luôn sẵn sàng khi cần thiết. Việc quản lý dữ liệu cần bao gồm cả việc làm sạch và chuẩn hóa dữ liệu để đảm bảo tính chính xác và nhất quán.
* Sử dụng công nghệ phù hợp: Cần lựa chọn các công cụ và công nghệ phù hợp để quản lý và lưu trữ dữ liệu. Các công nghệ này cần hỗ trợ việc phân tích dữ liệu và tạo ra các báo cáo, biểu đồ trực quan để giúp dễ dàng hiểu và sử dụng dữ liệu, đặc biệt giá thành, chi phí vận hành ở mức tối thiểu nhưng vẫn mang lại hiệu quả cao.
* Dữ liệu sẽ được quản lý và lưu trữ trên Excel nhằm đảm bảo tính bảo mật và dễ dàng truy cập, đồng thời cho phép thực hiện các truy vấn phân tích nhanh chóng. Excel cung cấp một môi trường lưu trữ dữ liệu ổn định, phù hợp cho các tập dữ liệu lớn với nhiều bảng và hàng trăm nghìn bản ghi. Hệ thống sẽ được thiết lập các quyền truy cập rõ ràng, chỉ những nhân viên có quyền mới có thể thao tác trên dữ liệu, điều này giúp bảo vệ tính bảo mật của dữ liệu.
* Dữ liệu cần được phân loại và chuẩn hóa theo từng bảng dữ liệu, cùng với việc thiết lập các khóa ngoại để đảm bảo tính toàn vẹn và liên kết giữa các bảng. Để quản lý dữ liệu hiệu quả, mỗi bảng sẽ có các chỉ mục phù hợp nhằm tăng tốc độ truy vấn. Tất cả dữ liệu sẽ được kiểm soát chặt chẽ qua các quyền truy cập và cơ chế bảo mật của hệ thống, đảm bảo chỉ những nhân viên có quyền truy cập mới có thể thao tác trên dữ liệu, từ đó giúp bảo vệ tính bảo mật của dữ liệu.

**Mục tiêu**

# Mục tiêu chính của việc thu thập dữ liệu là để phân tích tuyển dụng trong ngành công nghệ thông tin (CNTT) tại Việt Nam.

* Phạm vi: Việt Nam
* Đối tượng: Các vị trí công việc trong ngành CNTT như lập trình viên, quản lý dự án, chuyên gia an ninh mạng, kỹ sư dữ liệu, chuyên gia AI, và các vị trí liên quan khác.
* Nguồn dữ liệu: Thu thập dữ liệu ở các trang web bởi nguồn tin tuyển dụng ở các trang web có 1 lượng thông tin nhất định theo yêu cầu của trang web, nếu muốn đăng 1 bản tin tuyển dụng. Hơn nữa, sự đa dạng số lượng trang web để đăng các tin tuyển dụng là yếu tố
* Chính để nhóm quyết định thu thập dữ liệu từ trang web.

| Trang web | Thành viên phụ trách |
| --- | --- |
| Career Builder | Khải, Việt Anh |
| Vieclam24h | Hùng |
| TopCV | Thiện |
| Vietnamwork | Quyền |

*Bảng 1.2: Bảng thông tin trang web thu thập*

* Các trang web để thu thập dữ liệu bao gồm:*1.2Bảng thông tin trang web thu thập*
* Lý do sử dụng các trang web trên: Như chủ đề có phạm vi là Việt Nam nên nhóm sử dụng các trang web có các tin tuyển dụng của Việt Nam, có định dạng khá giống nhau nên lượng thông tin cung cấp cũng tương đồng giúp dễ dàng phần nào khi ghép các bản ghi thuộc các trang web khác nhau.

**Đánh giá tính khả thi**  
 - Năng lực (skill hiện có): Với kinh nghiệm và kiến thức sẵn có, tôi đã có thể thực hiện các công việc cơ bản trong dự án này. Hiện tại, tôi có các kỹ năng mạnh về SQL, Python, NumPy, Excel, Power BI, và kiến thức vững về toán xác suất thống kê. Những kỹ năng này giúp tôi dễ dàng trong việc truy vấn, xử lý dữ liệu lớn, phân tích định lượng và xây dựng báo cáo chi tiết về các xu hướng và hành vi khách hàng. Bên cạnh đó, khả năng kể chuyện qua dữ liệu là một thế mạnh, giúp tôi truyền tải các kết quả phân tích một cách rõ ràng, dễ hiểu qua các báo cáo trực quan. Power BI là công cụ tôi sử dụng để tạo các biểu đồ và dashboard chuyên nghiệp, giúp các bên liên quan có cái nhìn trực quan và toàn diện về dữ liệu.  
 - Năng lực (skill sẽ học thêm cho dự án): Để đạt được mục tiêu dài hạn của dự án và tối ưu hóa quá trình phân tích, tôi sẽ bổ sung kiến thức và kỹ năng trong một số lĩnh vực mới. Tôi dự định học thêm về thuật toán để xử lý dữ liệu nhanh chóng và trực quan hóa dữ liệu nâng cao trên Python. Bên cạnh đó, để thực hiện các yêu cầu dự báo và dự đoán, tôi sẽ nghiên cứu các phương pháp xử lí dữ liệu dạng văn bản. Matplotlib cũng là công cụ tôi sẽ học để tạo ra các biểu đồ chuyên sâu, linh hoạt và trực quan hơn, phù hợp cho việc trình bày dữ liệu trong các báo cáo chi tiết. Với việc kết hợp thêm các kỹ năng này, tôi tin tưởng rằng mình có đủ khả năng để đáp ứng yêu cầu và hoàn thành dự án một cách hiệu quả.

Tóm lại, với nền tảng kỹ năng hiện có và kế hoạch nâng cao kỹ năng cần thiết, tôi có đầy đủ năng lực để triển khai dự án này. Tôi sẽ phối hợp chặt chẽ với các bên liên quan để đảm bảo rằng các phân tích đáp ứng đúng mục tiêu và mang lại giá trị cao nhất từ dữ liệu.

## lập kế hoạch dự án

| **TT** | **hạng mục** | **bắt đầu** | **kết thúc** | **thành viên** | **kết quả** |
| --- | --- | --- | --- | --- | --- |
| 1 | Giới thiệu dự án | 4/11/2024 | 4/11/2024 | Khải | 100% |
| 1.1 | Giới thiệu | 4/11/2024 | 4/11/2024 | Khải | 100% |
| 1.2 | Yêu cầu | 4/11/2024 | 4/11/2024 | Quyền,Khải, Hùng | 100% |
| 1.3 | Lập kế hoạch dự án | 9/11/2024 | 12/11/2024 | Khải | 100% |
| 2 | Phân tích yêu cầu KH | 9/11/2024 | 11/11/2024 | Cả nhóm | 100% |
| 2.1 | Phân tích yêu cầu | 9/11/2024 | 11/11/2024 | Quyền | 100% |
| 2.2 | Câu chuyện dữ liệu | 9/11/2024 | 11/11/2024 | Thiện, Quyền | 100% |
| 2.3 | Kiến trúc hệ thống | 12/11/2024 | 13/11/2024 | Khải | 100% |
| 2.4 | Giải thích về bộ dữ liệu khách hàng | 12/11/2024 | 13/11/2024 | Cả nhóm | 100% |
| 3 | Làm sạch và chuyển đổi dữ liệu | 13/11/2024 | 13/11/2024 | Thiện,Hùng | 100% |
| 3.1 | Chuẩn bị dữ liệu | 13/11/2024 | 13/11/2024 | Thiện | 100% |
| 3.2 | Làm sạch dữ liệu | 13/11/2024 | 13/11/2024 | Thiện | 100% |
| 3.3 | Chuyển đổi dữ liệu | 13/11/2024 | 13/11/2024 | Khải | 100% |
| 4 | Xử lý dữ liệu | 13/11/2024 | 14/11/2024 | Khải | 100% |
| 4.1 | Chuẩn hóa dữ liệu | 13/11/2024 | 14/11/2024 | Hùng | 100% |
| 4.2 | Mô hình hóa dữ liệu | 13/11/2024 | 14/11/2024 | Hùng, Khải | 100% |
| 4.3 | Xử lý dữ liệu DAX | 13/11/2024 | 14/11/2024 | Thiện | 100% |
| 5 | Trực quan hóa dữ liệu | 15/11/2024 | 16/11/2024 | Cả nhóm | 100% |
| 5.1 | Các kĩ thuật trực quan hóa | 16/11/2024 | 16/11/2024 | Hùng | 100% |
| 5.2 | Các nguyên tắc trực quan hóa | 16/11/2024 | 17/11/2024 | Hùng | 100% |
| 5.3 | Trình bày cách thêm visual mới | 17/11/2024 | 18/11/2024 | Cả nhóm | 100% |
| 5.4 | Trình bày tạo các report cho dự án | 17/11/2024 | 19/11/2024 | Hùng | 100% |
| 6 | Xây dựng báo cáo | 20/11/2024 | 24/11/2024 | Cả nhóm | 100% |
| 6.1 | Dashboard và Report | 20/11/2024 | 24/11/2024 | Thiện | 100% |
| 6.2 | Xây dựng báo cáo | 21/11/2024 | 24/11/2024 | Thiện | 100% |
| 7 | Kết luận | 25/11/2024 | 26/11/2024 | Cả nhóm | 100% |
| 7.1 | Báo cáo | 25/11/2024 | 26/11/2024 | Cả nhóm | 100% |
| 7.2 | Khó khăn | 26/11/2024 | 27/11/2024 | Quyền | 100% |
| 7.3 | Thuận lợi | 26/11/2024 | 28/11/2024 | Quyền | 100% |
| 7.4 | Hướng phát triển | 27/11/2024 | 29/11/2024 | Quyền | 100% |
| 8 | Tổng kết | 29/11/2024 | 29/11/2024 | Khải | 100% |

*Bảng 1.3: Kế hoạch chi tiết*

# phân tích yêu cầu tuyển dụng

## phân tích yêu cầu

**Dữ liệu:**

* **Tổng quan:** Tổng số công việc, trung bình lương min, trung bình lương max, trung bình kinh nghiệm, số lượng công việc theo tháng, sự chênh lệch lương max và lương min.
* **Phân tích mức lương:** Số bản tin lương cứng, số bản tin lương cạnh tranh, trung bình lương min, trung bình lương max, lương min và lương max theo nhóm công việc, tỷ lệ lương cạnh tranh và lương trung bình, lương theo kinh nghiệm, lương trung bình theo top 10 thành phố.
* **Phân tích kinh nghiệm và cấp bậc:** Kinh nghiệm trung bình, kinh nghiệm của thực tập sinh, số công việc của thực tập sinh, phần trăm công việc của thực tập sinh, số lượng công việc của thực tập sinh, số lượng công việc theo cấp bậc, kinh nghiệm trung bình theo từng cấp bậc, top thành phố yêu cầu kinh nghiệm trung bình thấp nhất.
* **Phân tích ngành nghề và nhóm công việc:** Top ngành nghề có lương cao nhất, top nhóm công việc có lương cao nhất, top 10 ngành nghề theo số lượng công việc, nhóm công việc theo số lượng công việc.
* **Phân tích thành phố và ngoại ngữ:** Tỷ lệ việc làm theo ngoại ngữ, mức lương trung bình theo ngoại ngữ, top 5 thành phố theo số lượng công việc, top thành phố có lương cao nhất.

**Quản lý và lưu trữ:**

* Dữ liệu được lưu trữ trong file excel, csv.

**Công nghệ:**

* Sử dụng Power BI để trực quan hóa dữ liệu và tạo báo cáo.
* Sử dụng các kỹ thuật thống kê và học máy để phân tích chuyên sâu và dự đoán xu hướng.
* Sử dụng ngôn ngữ lập trình Python và các thư viện như Pandas, NumPy để xử lý và phân tích dữ liệu.
* Sử dụng các công cụ AI hỗ trợ lên kế hoạch, phân tích, tạo ảnh, video, …

**Quyết định sử dụng các công nghệ:**

* **Dữ liệu**: Cơ sở dữ liệu.
* **Quản lý và lưu trữ:**

Excel

Ngôn ngữ lập trình Python và các thư viện Pandas, NumPy.

* **Công nghệ:**

Công cụ trực quan hóa dữ liệu và tạo báo cáo Power BI.

Các kỹ thuật thống kê và học máy (clustering).

* AI: ChatGPT, Gemini, invideo, Aicolor…

**Giải thích:**

* **Cơ sở dữ liệu quan hệ:** Phù hợp để lưu trữ dữ liệu có cấu trúc và quan hệ giữa các bảng.
* **Excel:** Ngôn ngữ truy vấn dữ liệu mạnh mẽ, cho phép truy vấn và xử lý dữ liệu hiệu quả.
* **Python:** Ngôn ngữ lập trình phổ biến trong lĩnh vực khoa học dữ liệu, với các thư viện mạnh mẽ hỗ trợ phân tích dữ liệu.
* **Power BI:** Công cụ mạnh mẽ để trực quan hóa dữ liệu và tạo báo cáo tương tác.
* **Thống kê và học máy:** Giúp phân tích chuyên sâu và dự đoán xu hướng.

## câu chuyện dữ liệu

### đặt vấn đề

### Khởi đầu và vấn đề của Hùng

* Hùng, một sinh viên mới tốt nghiệp ngành Xử lý dữ liệu tại trường FPT Polytechnic, đang đối mặt với khó khăn trong việc tìm kiếm công việc phù hợp. Anh nhận ra rằng nhiều nhà tuyển dụng yêu cầu kinh nghiệm hoặc kỹ năng vượt quá khả năng hiện tại của mình.

### Phát hiện bản tin phân tích xu hướng tuyển dụng IT

* Trong lúc tìm kiếm cơ hội việc làm, Hùng đọc được một bản tin phân tích từ các nền tảng như TopCV, Vieclam24h, LinkedIn và Indeed. Bản tin cung cấp thông tin về 3.691 công việc với số lượng tuyển dụng cao nhất vào tháng 10 và thấp nhất vào tháng 8.

### Phân bố việc làm theo địa lý

* Bản tin chỉ ra Hà Nội là trung tâm tuyển dụng lớn nhất với 1.768 công việc, tiếp theo là thành phố Hồ Chí Minh (1.275 việc), cùng các tỉnh lân cận như Bình Dương, Đà Nẵng, và Hải Phòng. Hùng cảm thấy an tâm vì Hà Nội – nơi anh sống – là khu vực có cơ hội việc làm dồi dào.

### Mâu thuẫn trong mức lương và cơ hội ngành nghề

* Hùng nhận ra sự chênh lệch rõ ràng giữa mức lương các ngành: Bảo mật dẫn đầu với 48 triệu đồng, tiếp theo là Quản lý (47 triệu đồng). Trong khi đó, nhóm Phân tích và tư vấn khách hàng chỉ đạt mức thấp nhất (19 triệu đồng). Đối với chuyên ngành Data Engineer của Hùng, mức lương cao nhất là 43 triệu đồng, nhưng cơ hội lại khá ít với chỉ 43 vị trí và 12 vị trí dành cho thực tập sinh.

### Niềm hy vọng và kế hoạch của Hùng

* Mặc dù có những khó khăn, Hùng nhận ra tiềm năng của lĩnh vực IT khi mức lương trung bình dao động từ 16,3 triệu đến 25,2 triệu đồng. Anh quyết định học thêm các khóa kỹ năng, làm các dự án thực tế, và tận dụng thế mạnh chuyên ngành để sớm chinh phục thị trường việc làm IT đầy triển vọng.

### xác định câu chuyện

**Thông điệp từ dữ liệu**:  
Dữ liệu không chỉ mô tả thực trạng tuyển dụng mà còn nhấn mạnh tầm quan trọng của việc chuẩn bị kỹ năng phù hợp và khai thác cơ hội tại các lĩnh vực có tiềm năng phát triển cao như Data Engineer. Đối với những người mới tốt nghiệp như Hùng, việc chủ động trau dồi năng lực và tập trung vào ngành nghề có giá trị cao sẽ mở ra cánh cửa đến với sự nghiệp thành công.

**Mục tiêu**:

* Định hướng sự nghiệp cho sinh viên mới tốt nghiệp.
* Khuyến khích nâng cao kỹ năng cá nhân.
* Tận dụng dữ liệu để ra quyết định.
* Khai thác tiềm năng của ngành IT tại Việt Nam.

**Giải pháp**:

* Phát triển kỹ năng và tích lũy kinh nghiệm.
* Mở rộng cơ hội qua mạng lưới nghề nghiệp.
* Cá nhân hóa trải nghiệm mua sắm.
* Chuẩn bị tốt hồ sơ ứng tuyển.
* Tìm kiếm công việc phù hợp.

**Cách tiếp cận**:

* Hiểu rõ bối cảnh và nhu cầu thị trường.
* Phân tích dữ liệu thị trường.
* Xây dựng lộ trình nâng cao năng lực.

### xác định rõ đối tượng

**Đối tượng**:

**Sinh viên mới tốt nghiệp ngành IT**:

* **Nhu cầu**: Hiểu rõ thị trường lao động IT, yêu cầu kỹ năng và mức lương, đồng thời tìm được hướng đi phát triển sự nghiệp.
* **Lợi ích**: Trang bị kiến thức và kỹ năng để bước vào thị trường lao động đầy cạnh tranh.

**Người đang cân nhắc chuyển ngành sang IT**:

* **Nhu cầu**: Tìm hiểu về các ngành nghề trong IT, mức lương và cơ hội việc làm để đưa ra quyết định phù hợp.
* **Lợi ích**: Xác định lộ trình học tập và tích lũy kỹ năng để gia nhập ngành IT.

**Nhà tuyển dụng và tổ chức đào tạo**:

* **Nhu cầu**: Hiểu rõ xu hướng tuyển dụng, đặc điểm ứng viên tiềm năng, và nhu cầu đào tạo.
* **Lợi ích**: Xây dựng chương trình tuyển dụng và đào tạo hiệu quả, tập trung vào các kỹ năng thị trường đang cần.

**Những người quan tâm đến phân tích dữ liệu:**

* Nhu cầu: Học hỏi cách áp dụng dữ liệu để đưa ra các quyết định chiến lược.
* Lợi ích: Hiểu cách sử dụng dữ liệu để kể một câu chuyện ý nghĩa, từ đó đưa ra giải pháp cho các vấn đề thực tế.

**Phân tích đối tượng**:

**Sinh viên mới tốt nghiệp ngành IT**:

* Vừa hoàn thành chương trình học, chưa có nhiều kinh nghiệm thực tế.
* Đang trong giai đoạn tìm hiểu và chuẩn bị gia nhập thị trường lao động.

**Người muốn chuyển ngành sang IT**:

* Đã có kinh nghiệm trong lĩnh vực khác, muốn chuyển hướng sang ngành IT do tiềm năng phát triển và mức lương hấp dẫn.
* Có thể gặp rào cản về kỹ năng kỹ thuật hoặc thiếu kiến thức nền tảng.

**Nhà tuyển dụng và tổ chức đào tạo**:

* Nhà tuyển dụng: Các công ty công nghệ hoặc doanh nghiệp đang tìm kiếm nhân tài trong ngành IT.
* Tổ chức đào tạo: Các trung tâm hoặc trường học cung cấp các khóa học kỹ năng công nghệ.

**Người yêu thích phân tích dữ liệu**:

* Quan tâm đến việc ứng dụng phân tích dữ liệu vào giải quyết vấn đề thực tế.
* Có kỹ năng hoặc mong muốn học hỏi thêm về storytelling và data visualization.

### xác định câu chuyện chi tiết

**Bối cảnh**: Câu chuyện diễn ra trong thị trường lao động ngành IT tại Việt Nam, nơi nhu cầu tuyển dụng đang tăng trưởng mạnh mẽ nhưng yêu cầu khắt khe về kỹ năng và kinh nghiệm.

**Các bên liên quan**: Hùng (Nhân vật chính), nhà tuyển dụng, tổ chức đào tạo và người kể chuyện (Bản tin phân tích).

**Mạch truyện**:

**Bắt đầu**

* Hùng tốt nghiệp ngành Xử lý dữ liệu và đối mặt với áp lực tìm kiếm việc làm trong ngành IT.
* Anh nhận thấy thị trường lao động IT cạnh tranh, nhưng lại thiếu thông tin cụ thể về cơ hội và yêu cầu.

**Tiến triển**

* Hùng tình cờ đọc được một bản tin phân tích xu hướng tuyển dụng IT Việt Nam, cung cấp thông tin chi tiết về mức lương, cơ hội việc làm, và kỹ năng cần thiết.
* Anh nhận ra tiềm năng trong ngành Data Engineer với mức lương hấp dẫn (43 triệu đồng) và nhiều cơ hội tại Hà Nội.

**Xung đột**

* Hùng nhận ra mình thiếu kinh nghiệm thực tế và một số kỹ năng cần thiết để cạnh tranh trong thị trường lao động.
* Anh cảm thấy áp lực khi phải đáp ứng các yêu cầu cao từ nhà tuyển dụng.

**Cao trào**

* Hùng lập kế hoạch cụ thể để nâng cao kỹ năng: tham gia khóa học chuyên sâu, cải thiện hồ sơ ứng tuyển, và tìm kiếm các buổi hội thảo nghề nghiệp để mở rộng kết nối.

**Kết quả**

* Nhờ chuẩn bị kỹ lưỡng và sử dụng thông tin từ bản tin, Hùng trở nên tự tin hơn và sẵn sàng ứng tuyển vào các vị trí Data Engineer tại Hà Nội.

**Kết thúc**

* Câu chuyện khép lại với thông điệp tích cực: việc khai thác và phân tích dữ liệu đúng cách không chỉ giúp định hướng nghề nghiệp mà còn mang lại cơ hội phát triển lớn trong ngành IT.

### trình bày dữ liệu

| **Câu chuyện chi tiết** | **Biểu đồ** | **Mô tả** |
| --- | --- | --- |
| Mức độ phổ biến của các công việc IT theo khu vực. | Biểu đồ cột thể hiện số lượng việc làm IT ở các thành phố và tỉnh (Hà Nội, TP.HCM, Bình Dương, Đà Nẵng, Hải Phòng). | Biểu đồ này giúp Hùng và các nhà phân tích thấy được phân bổ công việc theo khu vực, giúp Hùng an tâm về cơ hội việc làm tại Hà Nội và có thể định hướng việc làm phù hợp. |
| Sự chênh lệch về mức lương giữa các ngành IT. | Biểu đồ cột thể hiện mức lương trung bình của các ngành IT (Bảo mật, Quản lý, Phân tích, Data Engineer, v.v.) | Biểu đồ này làm rõ sự chênh lệch mức lương giữa các ngành, giúp Hùng nhận ra các ngành có cơ hội kiếm được lương cao và từ đó cân nhắc điều chỉnh kế hoạch nghề nghiệp. |
| Mức lương trung bình và cơ hội việc làm trong ngành Data Engineer. | Biểu đồ tròn thể hiện tỷ lệ các ngành trong lĩnh vực IT với số lượng việc làm và mức lương tương ứng. | Biểu đồ này cung cấp cái nhìn rõ ràng về cơ hội nghề nghiệp và mức lương trong các ngành khác nhau, từ đó giúp Hùng hiểu rõ sự cạnh tranh trong ngành Data Engineer và quyết định nếu cần tìm thêm các ngành khác có nhiều cơ hội hơn. |
| Kế hoạch phát triển nghề nghiệp của Hùng. | Biểu đồ đường thể hiện sự thay đổi của mức lương trong các năm qua để thấy rõ xu hướng tăng trưởng lương trong ngành IT. | Biểu đồ này giúp Hùng nhận thức rõ về xu hướng phát triển nghề nghiệp trong ngành IT, từ đó lên kế hoạch học thêm kỹ năng và tham gia vào các dự án thực tế để nâng cao năng lực và đáp ứng được yêu cầu công việc. |

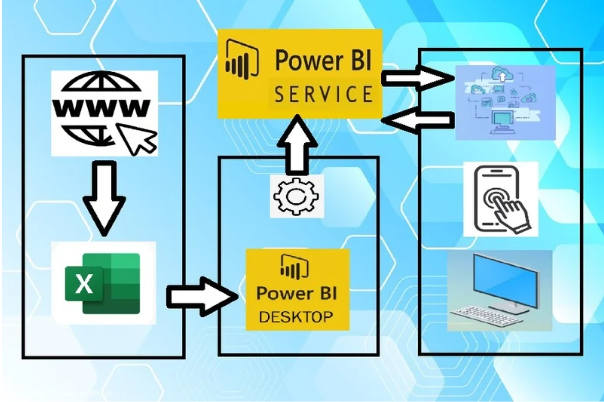
*Bảng 2.1: Trình bày dữ liệu*

### những điều cần lưu ý

* **Rõ ràng**: Dữ liệu và thông tin cần dễ hiểu.
* **Hấp dẫn**: Sử dụng hình ảnh, màu sắc, bố cục hợp lý.
* **Xúc tích**: Tránh quá tải thông tin.
* **Kết nối**: Liên kết các thông tin, tạo thành mạch truyện.
* **Hành động**: Đưa ra khuyến nghị, giải pháp.

## kiến trúc hệ thống

### kiến trúc

*Hình 2.1: Mô hình kiến trúc hệ thống*

**Excel:**

* Dữ liệu sẽ được lưu trữ trên Excel, bao gồm thông tin gốc và thông tin chi tiêu cuối cùng. Bảng dữ liệu Excel này lưu trữ tất cả các thông tin ban đầu trước khi qua bất kỳ quá trình xử lý nào và các thông tin chi tiêu đã được xử lý và tiền xử lý.

**Nguồn dữ liệu:**

* Dữ liệu được crawl từ các trang web như : topcv, vietnamwork, vieclam24h, career builder.

**Power BI Desktop:**

* Power BI Desktop kết nối với file Excel để lấy dữ liệu nhằm mục đích phân tích và trực quan hóa.
* Dữ liệu này sau đó được sử dụng để tạo báo cáo và bảng thông tin.

**Power BI Service:**

* Các báo cáo và bảng thông tin được tạo trong Power BI Desktop sẽ được xuất bản lên Power BI Service.
* Dịch vụ Power BI cho phép làm mới dữ liệu, chia sẻ và cộng tác trên các báo cáo và bảng thông tin.

**Người sử dụng/ Nhân viên:** Nhân viên truy cập Dịch vụ Power BI bằng nhiều thiết bị khác nhau:

* Máy tính để bàn
* Máy tính bảng
* Điện thoại di động

**Luồng dữ liệu có thể được tóm tắt như sau:**

1. Dữ liệu được thu thập từ các trang web và được chuyển vào file excel .
2. Dữ liệu đã được xử lý trong file Excel.
3. Power BI Desktop truy cập file Excel để tạo hình ảnh trực quan và báo cáo.
4. Những hình ảnh trực quan và báo cáo này được xuất bản trên Dịch vụ Power BI.
5. Nhân viên truy cập dịch vụ Power BI bằng nhiều thiết bị khác nhau để xem và tương tác với các báo cáo và bảng thông tin.

### giải thích

**Tổng quan thiết kế**

**Nguồn dữ liệu:** Dữ liệu được thu thập từ các trang web và được chuyển vào file excel

**Excel:**

* Dữ liệu sẽ được lưu trữ trên Excel, bao gồm thông tin gốc và thông tin chi tiêu cuối cùng. Bảng dữ liệu Excel này lưu trữ tất cả các thông tin ban đầu trước khi qua bất kỳ quá trình xử lý nào và các thông tin chi tiêu đã được xử lý và tiền xử lý.

**Công cụ phân tích và trực quan hóa:**

* **Power BI Desktop:** Công cụ này kết nối với Excel để lấy dữ liệu đã được xử lý. Power BI Desktop được sử dụng để tạo các biểu đồ, báo cáo và bảng điều khiển.
* **Power BI Service:** Sau khi tạo ra các báo cáo và bảng điều khiển trong Power BI Desktop, chúng được xuất bản lên Power BI Service. Power BI Service cung cấp một nền tảng trực tuyến để lưu trữ, quản lý và chia sẻ các báo cáo này.

**Người dùng cuối (Employees):** Nhân viên truy cập Power BI Service qua các thiết bị như máy tính bàn, máy tính bảng và điện thoại di động. Họ sử dụng các báo cáo và bảng điều khiển để thu thập thông tin và đưa ra các quyết định dựa trên dữ liệu.

**Quy trình và luồng dữ liệu**

**Thu thập dữ liệu:** Dữ liệu được thu thập từ các trang web công việc tuyển dụng tại Việt Nam

**Xử lý và chuẩn bị dữ liệu:** Dữ liệu từ các trang web được xử lý và làm sạch sau đó lưu vào file excel.

**Lưu trữ dữ liệu đã xử lý:** Dữ liệu đã được xử lý trong file Excel.

**Phân tích và tạo báo cáo:** Power BI Desktop kết nối với file Excel để lấy dữ liệu và tạo các biểu đồ, báo cáo, bảng điều khiển. Các báo cáo này sau đó được xuất bản lên Power BI Service.

**Truy cập và sử dụng dữ liệu:** Nhân viên sử dụng các thiết bị khác nhau để truy cập Power BI Service, nơi họ có thể xem và tương tác với các báo cáo và bảng điều khiển.

**Lợi ích của thiết kế hệ thống**

**Bảo mật dữ liệu:** Việc sử dụng Excel được bảo vệ an toàn và chỉ có những người được ủy quyền mới có thể truy cập và chỉnh sửa.

**Quản lý dữ liệu hiệu quả:** Sử dụng các cơ sở dữ liệu riêng biệt cho từng giai đoạn xử lý giúp quản lý dữ liệu một cách hiệu quả và có tổ chức.

**Trực quan hóa dữ liệu mạnh mẽ:** Power BI Desktop và Power BI Service cung cấp các công cụ mạnh mẽ để tạo ra các biểu đồ, báo cáo và bảng điều khiển, giúp người dùng dễ dàng hiểu và sử dụng dữ liệu.

**Truy cập linh hoạt:** Nhân viên có thể truy cập dữ liệu từ nhiều thiết bị khác nhau, giúp họ có thể làm việc mọi lúc, mọi nơi và ra quyết định nhanh chóng.

## giải thích về bộ dữ liệu tuyển dụng

### các khái niệm

**Dữ liệu thô (Raw Data):** Dữ liệu chưa qua xử lý, được nhập trực tiếp từ các nguồn như file CSV, hệ thống giao dịch, v.v.

**Làm sạch dữ liệu (Data Cleaning):** Quá trình loại bỏ hoặc sửa chữa các lỗi, dữ liệu không hợp lệ, và các giá trị thiếu trong dữ liệu.

**Chuẩn hóa dữ liệu (Data Normalization):** Quá trình tổ chức dữ liệu để giảm sự dư thừa và cải thiện tính toàn vẹn của dữ liệu.

### các trường dữ liệu

| **STT** | **Trường dữ liệu** | **Mô tả** |
| --- | --- | --- |
| 1 | Tên công việc | Tên vị trí tuyển dụng |
| 2 | Tên công ty | Tên doanh nghiệp hoặc tổ chức tuyển dụng |
| 3 | Ngày cập nhật | Ngày gần nhất thông tin công việc được cập nhật trên hệ thống |
| 4 | Ngành nghề | Lĩnh vực hoạt động hoặc ngành mà công việc thuộc về |
| 5 | Lương max | Mức lương đề xuất hoặc thỏa thuận, có thể là một khoảng hoặc cụ thể |
| 6 | Lương min | Mức lương đề xuất hoặc thỏa thuận, có thể là một khoảng hoặc cụ thể |
| 7 | Lương cạnh tranh | Mức lương đề xuất hoặc thỏa thuận, có thể là một khoảng hoặc cụ thể |
| 8 | Kinh nghiệm | Số năm kinh nghiệm yêu cầu |
| 9 | Cấp bậc | Vị trí trong hệ thống cấp bậc của công ty |
| 10 | Phúc lợi | Các quyền lợi dành cho nhân viên |
| 11 | Thành phố | Nơi làm việc hoặc văn phòng của công ty |
| 12 | Mô tả công việc | Nội dung chi tiết về trách nhiệm và nhiệm vụ mà người ứng tuyển sẽ đảm nhận |
| 13 | Yêu cầu công việc | Các điều kiện cần có từ ứng viên |
| 14 | Nhóm công việc | Phân nhóm công việc dựa theo mô tả công việc |
| 15 | Ngôn ngữ | Ngôn ngữ có sử dụng tiếng anh không. |
| 16 | Trình độ học vấn | Trình độ giáo dục tối thiểu mà công việc yêu cầu |

*Hình 2.2: các trường dữ liệu*

# làm sạch và chuyển đổi dữ liệu

## chuẩn bị dữ liệu

### giải pháp lưu trữ dữ liệu

Lưu trữ dữ liệu là một phần quan trọng của bất kỳ hệ thống quản lý dữ liệu nào. Việc chọn lựa giải pháp lưu trữ phù hợp phụ thuộc vào nhiều yếu tố như quy mô dữ liệu, yêu cầu về bảo mật, khả năng mở rộng, chi phí và khả năng truy cập. Hiện nay, có hai giải pháp lưu trữ chính: giải pháp nền tảng đám mây (Cloud-based solution) và các ứng dụng tại chỗ (On-premise).

**So sánh cloud-based solution and on-premise**

**Giải pháp nền tảng đám mây (Cloud-based solution)**

Ưu điểm:

* Khả năng mở rộng: Dễ dàng mở rộng hoặc thu nhỏ tài nguyên dựa trên nhu cầu thực tế mà không cần đầu tư cơ sở hạ tầng mới.
* Chi phí: Trả phí theo mức sử dụng (pay-as-you-go), không cần chi phí đầu tư ban đầu lớn cho phần cứng và phần mềm.
* Truy cập từ xa: Có thể truy cập dữ liệu từ bất kỳ đâu có kết nối internet.
* Sao lưu và khôi phục: Các nhà cung cấp dịch vụ đám mây thường cung cấp các giải pháp sao lưu và khôi phục dữ liệu tự động.
* Bảo mật và tuân thủ: Các nhà cung cấp dịch vụ đám mây thường có các tiêu chuẩn bảo mật và tuân thủ nghiêm ngặt.

Nhược điểm:

* Chi phí dài hạn: Chi phí có thể tăng lên đáng kể theo thời gian nếu nhu cầu sử dụng tài nguyên lớn.
* Quản lý dữ liệu: Phụ thuộc vào nhà cung cấp dịch vụ trong việc quản lý và bảo mật dữ liệu.
* Tốc độ truy cập: Tốc độ truy cập dữ liệu có thể bị ảnh hưởng bởi chất lượng kết nối internet.

**Các ứng dụng tại chỗ (on-premise)**

Ưu điểm:

* Kiểm soát hoàn toàn: Doanh nghiệp có toàn quyền kiểm soát cơ sở hạ tầng và dữ liệu của mình.
* Bảo mật: Dữ liệu được lưu trữ tại chỗ, giảm thiểu rủi ro bị truy cập trái phép từ bên ngoài.
* Tốc độ truy cập: Tốc độ truy cập dữ liệu có thể nhanh hơn do không phụ thuộc vào kết nối internet.

Nhược điểm:

* Chi phí đầu tư ban đầu: Chi phí đầu tư ban đầu cho phần cứng, phần mềm, và cơ sở hạ tầng là rất lớn.
* Khả năng mở rộng: Việc mở rộng cơ sở hạ tầng đòi hỏi thêm chi phí và thời gian.
* Bảo trì và nâng cấp: Doanh nghiệp phải tự chịu trách nhiệm về bảo trì, nâng cấp và quản lý cơ sở hạ tầng.
* Sao lưu và khôi phục: Cần có kế hoạch sao lưu và khôi phục dữ liệu chi tiết để tránh mất mát dữ liệu.

**Quyết định và lý do**

Lựa chọn: giải pháp ứng dụng tại chỗ (on-premise) sử dụng Excel.

Lý do:

* Bảo vệ tệp bằng mật khẩu**:** Excel cho phép người dùng bảo vệ toàn bộ tệp bằng mật khẩu, ngăn chặn truy cập trái phép vào nội dung của tài liệu. Khi mật khẩu được thiết lập, người dùng chỉ có thể mở tệp sau khi nhập đúng mật khẩu, giúp bảo vệ thông tin nhạy cảm khỏi việc bị lộ hoặc chỉnh sửa.
* Bảo vệ các trang tính và ô: Excel cung cấp khả năng bảo vệ các trang tính hoặc các ô cụ thể trong một tệp, cho phép chỉ những người có quyền truy cập mới có thể chỉnh sửa dữ liệu trong khi những phần khác của tài liệu vẫn có thể được xem. Điều này hữu ích khi có nhiều người làm việc trên cùng một tệp và bạn muốn bảo vệ các phần quan trọng của dữ liệu.
* Mã hóa tệp Excel: Khi bạn bảo vệ tệp Excel bằng mật khẩu, tệp sẽ được mã hóa. Điều này có nghĩa là chỉ có những người biết mật khẩu mới có thể mở và đọc nội dung trong tệp, ngay cả khi tệp bị sao chép và chia sẻ trái phép. Mã hóa giúp bảo vệ dữ liệu khi lưu trữ hoặc chia sẻ qua email hoặc các phương tiện khác.
* Theo dõi và ghi lại thay đổi: Excel cho phép bạn theo dõi và ghi lại mọi thay đổi được thực hiện trong tệp qua tính năng ‘Track Changes’ (Theo dõi thay đổi). Điều này rất hữu ích khi làm việc nhóm và giúp bạn biết ai đã thay đổi gì, khi nào và lý do thay đổi. Điều này giúp kiểm soát và quản lý các chỉnh sửa không mong muốn.
* Hạn chế quyền truy cập qua IRM (Information Rights Management): Thông qua tính năng IRM (Quản lý quyền thông tin), Excel cho phép bạn giới hạn quyền truy cập vào tài liệu và quy định các quyền hạn cụ thể như xem, chỉnh sửa, sao chép hoặc in tệp. Điều này giúp bạn bảo vệ dữ liệu khỏi việc bị sao chép hoặc sửa đổi mà không có sự cho phép.
* Khôi phục lịch sử phiên bản: Excel tích hợp tính năng Version History (Lịch sử phiên bản), cho phép bạn khôi phục lại các phiên bản trước của tệp nếu có sự cố hoặc thay đổi không mong muốn. Điều này đảm bảo rằng nếu dữ liệu bị thay đổi hoặc xóa nhầm, bạn có thể khôi phục lại dữ liệu trước đó.
* Bảo mật khi chia sẻ qua OneDrive hoặc SharePoint: Khi lưu trữ tệp Excel trên OneDrive hoặc SharePoint, bạn có thể kiểm soát quyền truy cập và chia sẻ tệp với người khác một cách an toàn. Bạn có thể đặt quyền truy cập theo cấp độ (chỉ xem, chỉnh sửa, v.v.) và theo dõi ai đã truy cập hoặc thay đổi tệp. Excel cũng hỗ trợ các tính năng bảo mật như xác thực hai yếu tố khi truy cập tài liệu trên đám mây.
* Bảo vệ dữ liệu với VBA (Visual Basic for Applications): VBA cho phép bạn tự động hóa nhiều tác vụ trong Excel và bảo vệ mã (macro) khỏi việc bị thay đổi hoặc truy cập trái phép. Bạn có thể sử dụng VBA để bảo vệ các quy trình tự động và dữ liệu quan trọng khỏi bị chỉnh sửa ngoài ý muốn.
* Giới hạn quyền chỉnh sửa qua quyền người dùng: Khi làm việc trong môi trường doanh nghiệp hoặc nhóm, Excel có thể thiết lập quyền hạn cho từng người dùng thông qua các cài đặt như ‘Allow Users to Edit Ranges’ (Cho phép người dùng chỉnh sửa các vùng dữ liệu). Điều này giúp bạn chỉ định những người có quyền chỉnh sửa các phần dữ liệu nhất định, bảo vệ các phần còn lại khỏi sự thay đổi.
* Chống lại mã độc và phần mềm độc hại: Excel cung cấp các tính năng bảo vệ để ngừng mã độc hoặc phần mềm độc hại có thể xâm nhập vào tệp, chẳng hạn như tính năng ‘Protected View’ (Chế độ bảo vệ) khi mở các tệp từ nguồn không xác định. Điều này giúp ngừng các chương trình nguy hiểm chạy trong khi bạn đang làm việc với tài liệu.

Tóm lại, Excel mang lại nhiều lợi ích bảo mật khi lưu trữ dữ liệu, từ việc bảo vệ tệp bằng mật khẩu cho đến các tính năng bảo mật nâng cao như mã hóa, IRM, và theo dõi thay đổi. Những tính năng này không chỉ giúp bảo vệ dữ liệu khỏi truy cập trái phép mà còn giúp theo dõi và quản lý các thay đổi trong tài liệu. Bằng cách sử dụng các công cụ bảo mật có sẵn, bạn có thể đảm bảo rằng dữ liệu của mình được bảo vệ an toàn và chỉ có những người được ủy quyền mới có thể truy cập và chỉnh sửa.

### giải pháp phân bố dữ liệu

Khi lưu trữ dữ liệu tuyển dụng IT dưới dạng Excel, bạn có thể phân bố và tổ chức dữ liệu theo các bước sau để đảm bảo dễ dàng phân tích và truy xuất thông tin.

**Tạo các sheet riêng biệt cho các nhóm dữ liệu**

**Sheet 1: Tổng quan dữ liệu tuyển dụng**

* Các trường: Mã tin tuyển dụng, Ngành nghề, Mức lương, Kinh nghiệm, Yêu cầu kỹ năng, Địa điểm, Mô tả công việc, Ngày đăng tin, v.v.

**Sheet 2: Phân bổ theo ngành nghề**

* Các trường: Mã tin tuyển dụng, Ngành nghề, Số lượng công việc, Mức lương trung bình, Yêu cầu kỹ năng chủ yếu, v.v.

**Sheet 3: Phân bổ theo địa lý**

* Các trường: Mã tin tuyển dụng, Địa điểm, Số lượng công việc theo khu vực, Mức lương trung bình theo khu vực.

**Sheet 4: Phân bổ theo mức lương**

* Các trường: Mã tin tuyển dụng, Mức lương, Ngành nghề, Số lượng công việc trong từng mức lương (ví dụ, <20 triệu, 20-30 triệu, 30 triệu trở lên).

**Sheet 5: Phân bổ theo kinh nghiệm**

* Các trường: Mã tin tuyển dụng, Kinh nghiệm yêu cầu, Số lượng công việc cho từng mức kinh nghiệm (ví dụ, "0-1 năm", "1-3 năm", "3-5 năm", "trên 5 năm").

**Tổ chức dữ liệu trong các sheet**

* **Cột dữ liệu chính**: Sử dụng các cột như "Mã tin tuyển dụng", "Ngành nghề", "Địa điểm", "Mức lương", "Kinh nghiệm", "Yêu cầu kỹ năng" để nhóm thông tin quan trọng lại với nhau.
* **Dữ liệu phân loại**: Sử dụng các cột bổ sung như "Mức lương thấp", "Mức lương cao", "Kinh nghiệm yêu cầu", "Loại công việc" để phân loại các công việc theo các nhóm.
* **Tạo cột phụ trợ cho phân tích**: Các cột như "Chênh lệch lương", "Sự phổ biến của ngành", "Tần suất kỹ năng yêu cầu" sẽ giúp bạn dễ dàng phân tích và tạo báo cáo

#### Ý nghĩa việc phân bố dữ liệu

Phân bố dữ liệu là quá trình tổ chức và quản lý dữ liệu sao cho dữ liệu được lưu trữ trên nhiều vị trí khác nhau, có thể là trên các máy chủ khác nhau trong cùng một hệ thống hoặc trên các hệ thống khác nhau. Việc phân bố dữ liệu có ý nghĩa quan trọng trong nhiều khía cạnh của quản lý dữ liệu, từ việc cải thiện hiệu suất truy cập, tăng cường tính sẵn sàng và độ tin cậy, đến việc đảm bảo an toàn và bảo mật dữ liệu.

**Các lợi ích của việc phân bố dữ liệu:**

1. Cải thiện hiệu suất truy cập dữ liệu:

* Giảm tải trên máy chủ: Phân bố dữ liệu giúp phân tán tải công việc giữa các máy chủ khác nhau, tránh tình trạng quá tải trên một máy chủ đơn lẻ. Điều này cải thiện hiệu suất truy cập dữ liệu và đảm bảo rằng hệ thống có thể phục vụ nhiều người dùng đồng thời mà không bị chậm trễ.
* Tăng tốc độ truy cập: Dữ liệu được phân bố trên nhiều vị trí có thể giúp giảm độ trễ truy cập, đặc biệt là khi các máy chủ được đặt gần với người dùng cuối hoặc các ứng dụng truy cập dữ liệu.

1. Tăng cường tính sẵn sàng và độ tin cậy:

* Khả năng dự phòng: Khi dữ liệu được sao chép và lưu trữ trên nhiều máy chủ, hệ thống có khả năng tiếp tục hoạt động ngay cả khi một hoặc nhiều máy chủ gặp sự cố. Điều này làm tăng tính sẵn sàng của hệ thống và giảm thiểu thời gian ngừng hoạt động.
* Khả năng khôi phục: Dữ liệu phân bố giúp dễ dàng khôi phục lại dữ liệu trong trường hợp có sự cố hoặc mất mát dữ liệu. Các bản sao dự phòng có thể được sử dụng để khôi phục hệ thống một cách nhanh chóng và hiệu quả.

1. Cải thiện khả năng mở rộng:

* Mở rộng tài nguyên dễ dàng: Phân bố dữ liệu cho phép hệ thống mở rộng dễ dàng bằng cách thêm các máy chủ mới mà không cần thay đổi cấu trúc cơ bản của hệ thống. Điều này giúp hệ thống có thể phát triển và đáp ứng nhu cầu ngày càng tăng của người dùng.
* Phân tán dữ liệu theo khu vực: Hệ thống có thể phân bố dữ liệu theo khu vực địa lý để phục vụ người dùng ở các vị trí khác nhau một cách hiệu quả hơn. Ví dụ, dữ liệu có thể được lưu trữ ở các máy chủ gần với người dùng khu vực đó để giảm độ trễ và tăng tốc độ truy cập.

1. Đảm bảo an toàn và bảo mật dữ liệu:

* Bảo vệ dữ liệu trước các mối đe dọa: Phân bố dữ liệu trên nhiều máy chủ giúp bảo vệ dữ liệu trước các mối đe dọa như tấn công mạng, thiên tai, hoặc lỗi phần cứng. Dữ liệu có thể được sao lưu và bảo mật tại nhiều vị trí khác nhau.
* Tuân thủ quy định bảo mật: Phân bố dữ liệu theo khu vực địa lý cũng giúp tuân thủ các quy định bảo mật và bảo vệ dữ liệu của từng quốc gia hoặc khu vực.

1. Tối ưu hóa chi phí:

* Sử dụng tài nguyên hiệu quả: Phân bố dữ liệu cho phép sử dụng hiệu quả các tài nguyên máy chủ hiện có, giúp tối ưu hóa chi phí vận hành và bảo trì hệ thống.
* Giảm chi phí đầu tư: Bằng cách sử dụng các giải pháp đám mây và dịch vụ nhân bản dữ liệu, doanh nghiệp có thể giảm chi phí đầu tư vào phần cứng và cơ sở hạ tầng, chỉ trả phí dựa trên mức sử dụng thực tế.

**Ứng dụng trong thực tế:**

1. Hệ thống thương mại điện tử

Các hệ thống thương mại điện tử thường phải xử lý lượng lớn dữ liệu giao dịch và truy cập từ nhiều người dùng trên toàn thế giới. Việc phân bố dữ liệu giúp cải thiện tốc độ truy cập, đảm bảo tính sẵn sàng của hệ thống, và bảo vệ dữ liệu người dùng.

1. Ứng dụng tài chính

Các ứng dụng tài chính cần đảm bảo tính nhất quán và an toàn của dữ liệu giao dịch. Phân bố dữ liệu giúp cải thiện hiệu suất và đảm bảo rằng dữ liệu luôn được bảo vệ và có thể khôi phục nhanh chóng trong trường hợp xảy ra sự cố.

1. Dịch vụ trực tuyến

Các dịch vụ trực tuyến như mạng xã hội, dịch vụ truyền thông, và các nền tảng nội dung số cần xử lý và lưu trữ lượng lớn dữ liệu người dùng. Phân bố dữ liệu giúp hệ thống mở rộng dễ dàng và đáp ứng nhu cầu truy cập cao từ người dùng.

#### Trình bày cách phân bố dữ liệu

Việc phân bố dữ liệu trong dự án mang lại nhiều lợi ích như cải thiện hiệu suất, tăng cường tính sẵn sàng, khả năng mở rộng, và bảo mật dữ liệu. Bằng cách thiết lập hệ thống nhân bản, phân bố dữ liệu theo khu vực địa lý, phân bố tải công việc và thiết lập sao lưu định kỳ, dự án có thể đảm bảo rằng dữ liệu luôn sẵn sàng và an toàn, đáp ứng tốt nhu cầu của người dùng cuối và giúp doanh nghiệp đưa ra các quyết định dựa trên dữ liệu một cách hiệu quả.

Trong câu chuyện của Hùng, dữ liệu về tuyển dụng IT có thể được phân bố và tổ chức để phục vụ cho việc phân tích sâu sắc các yếu tố như ngành nghề, mức lương, yêu cầu kỹ năng, khu vực địa lý, và mức kinh nghiệm. Dưới đây là cách phân bố dữ liệu theo các tiêu chí quan trọng từ câu chuyện của Hùng:

1. Sheet 1: Tổng quan dữ liệu tuyển dụng

* Đây là sheet chính tổng hợp tất cả các tin tuyển dụng mà Hùng đã tìm thấy trong bản tin phân tích xu hướng tuyển dụng IT.
* Các trường dữ liệu chính:
* Mã tin tuyển dụng: Để nhận diện từng tin tuyển dụng riêng biệt.
* Ngành nghề: Các ngành như Bảo mật, Quản lý, Data Engineering, Phân tích, v.v.
* Mức lương: Định dạng như "x triệu VND" hoặc "USD", với các mức lương từ thấp đến cao.
* Kinh nghiệm yêu cầu: Cung cấp mức kinh nghiệm yêu cầu cho từng công việc (0-1 năm, 1-3 năm, v.v.).
* Địa điểm: Các khu vực tuyển dụng như Hà Nội, TP.HCM, Đà Nẵng.
* Kỹ năng yêu cầu: Các kỹ năng cần có, như Python, SQL, Machine Learning,
* Mô tả công việc: Tóm tắt công việc mà nhà tuyển dụng yêu cầu.
* Ngày đăng tin: Thời điểm đăng tuyển dụng.

2. Sheet 2: Phân bổ theo ngành nghề

* Phân tích tổng số công việc và mức lương trong từng ngành nghề.
* Các trường dữ liệu:
* Ngành nghề: Phân loại theo các nhóm công việc như Bảo mật, Quản lý, Data Engineer, v.v.
* Số lượng công việc: Tổng số công việc theo từng ngành.
* Mức lương trung bình: Trung bình mức lương trong từng ngành (có thể chia theo cấp bậc).
* Yêu cầu kỹ năng chính: Các kỹ năng thường xuyên xuất hiện trong từng ngành.

3. Sheet 3: Phân bổ theo địa lý

* Tổ chức và phân tích dữ liệu theo các khu vực tuyển dụng để Hùng có cái nhìn rõ ràng về các cơ hội công việc tại các thành phố khác nhau.
* Các trường dữ liệu:
* Địa điểm: Các thành phố, tỉnh thành nơi có nhu cầu tuyển dụng cao (Hà Nội, TP.HCM, Đà Nẵng, v.v.).
* Số lượng công việc theo khu vực: Tổng số công việc được đăng tuyển theo từng khu vực.
* Mức lương trung bình theo khu vực: Mức lương trung bình của các công việc tại khu vực đó.
* Các kỹ năng yêu cầu theo khu vực: Tóm tắt các kỹ năng phổ biến trong từng khu vực.

4. Sheet 4: Phân bổ theo mức lương

* Phân tích sự phân bố của các mức lương khác nhau và sự liên quan của chúng với các yếu tố như ngành nghề, kinh nghiệm, và kỹ năng yêu cầu.
* Các trường dữ liệu:
* Mức lương: Các mức lương cụ thể như dưới 20 triệu, 20-30 triệu, 30 triệu trở lên.
* Số lượng công việc trong từng mức lương: Tổng số công việc có mức lương tương ứng.
* Ngành nghề: Ngành nghề có mức lương cao nhất và thấp nhất.
* Kinh nghiệm yêu cầu: Mức kinh nghiệm cho từng mức lương.

5. Sheet 5: Phân bổ theo kinh nghiệm yêu cầu

* Phân tích yêu cầu kinh nghiệm của các công việc tuyển dụng để giúp Hùng biết được thị trường đang yêu cầu những mức kinh nghiệm nào.
* Các trường dữ liệu:
* Kinh nghiệm yêu cầu: Các mức kinh nghiệm như 0-1 năm, 1-3 năm, 3-5 năm, trên 5 năm.
* Số lượng công việc: Số lượng công việc theo từng mức kinh nghiệm.
* Mức lương trung bình theo kinh nghiệm: Mức lương cho từng mức kinh nghiệm yêu cầu.
* Ngành nghề: Các ngành nghề yêu cầu mức kinh nghiệm tương ứng.

## làm sạch dữ liệu

### các vấn đề ảnh hưởng tới dữ liệu

**3.2.1.1** **Các vấn đề ảnh hưởng**

Dữ liệu là tài sản quan trọng của bất kỳ hệ thống nào. Tuy nhiên, có nhiều vấn đề có thể ảnh hưởng tới chất lượng và tính toàn vẹn của dữ liệu. Các vấn đề này có thể đến từ nhiều nguồn khác nhau và nếu không được xử lý kịp thời, chúng có thể gây ra hậu quả nghiêm trọng.

**Dữ liệu thiếu hoặc không chính xác:** Dữ liệu thiếu hoặc không chính xác có thể gây ra sai lệch trong phân tích và đưa ra quyết định.

Ví dụ: Trường “Mô tả công việc” khi thu thập từ các trang web bằng code có thể không có.

**Dữ liệu trùng lặp:** Dữ liệu trùng lặp làm tăng kích thước cơ sở dữ liệu không cần thiết và có thể dẫn đến các phân tích không chính xác.

Ví dụ: Các bản tin có thể trùng lặp, hoặc do người đăng một bản tin trong các ngày gần nhau.

**Dữ liệu ngoại lai (Outliers):** Các giá trị ngoại lai có thể làm méo mó các kết quả phân tích.

Ví dụ: Lương của một số công việc có mức lương cao bất thường.

**Vấn đề về bảo mật dữ liệu:** Dữ liệu nhạy cảm có thể bị truy cập trái phép nếu không được bảo mật đúng cách.

**Mất mát dữ liệu:** Dữ liệu có thể bị mất mát do lỗi phần cứng, lỗi phần mềm, hoặc do thảm họa tự nhiên.

**Đồng bộ hóa dữ liệu:** Dữ liệu không được đồng bộ hóa đúng cách giữa các hệ thống có thể dẫn đến sự không nhất quán.

**3.2.1.2** **Vấn đề đang tồn tại trong dự án**

Việc nhận diện và xử lý các vấn đề ảnh hưởng đến dữ liệu là rất quan trọng để đảm bảo chất lượng và tính toàn vẹn của dữ liệu. Trong phạm vi dự án này, việc làm sạch và chuẩn hóa dữ liệu, xử lý dữ liệu trùng lặp và ngoại lai, đồng bộ hóa dữ liệu đúng cách, và tối ưu hóa hiệu suất là các yếu tố then chốt để thành công.

### các tiêu chí đánh giá chất lượng dữ liệu

#### 3.2.2.1 Các tiêu chí

**Độ chính xác (Accuracy):** Dữ liệu phải phản ánh đúng thực tế mà nó đại diện. Các giá trị dữ liệu phải chính xác và không có sai sót.

Ví dụ: Mức lương của một số công việc phải phản ánh đúng thực tế

**Tính toàn vẹn (Integrity):** Dữ liệu phải nhất quán và không bị hỏng. Tính toàn vẹn đảm bảo rằng các quan hệ giữa các thực thể trong cơ sở dữ liệu được duy trì.

Ví dụ: Mỗi một công việc có một mã công việc.

**Tính nhất quán (Consistency):** Dữ liệu phải nhất quán trong toàn bộ cơ sở dữ liệu. Các giá trị tương tự phải được biểu diễn một cách thống nhất.

Ví dụ: Ngày tháng trong cột ngày cập nhật phải có cùng định dạng là dd/mm/yyyy.

**Tính dễ hiểu (Understandability):** Dữ liệu phải dễ hiểu và dễ sử dụng. Các trường dữ liệu và giá trị phải rõ ràng và có ý nghĩa.

**Tính truy xuất (Accessibility):** Dữ liệu phải dễ dàng truy cập và sử dụng bởi các bên liên quan. Điều này bao gồm cả việc đảm bảo dữ liệu có thể được truy cập một cách an toàn và bảo mật.

Ví dụ: Dữ liệu từ các bảng phải có thể truy cập từ Power BI và các công cụ phân tích khác một cách nhanh chóng và bảo mật.

#### 3.2.2.2 Tiêu chí áp dụng trong dự án

**Độ chính xác (Accuracy):**

* Lý do: Đảm bảo các giá trị dữ liệu ngày tháng phải phù hợp, mức lương phải thực tế.
* Cách thực hiện: Kiểm tra và xác minh dữ liệu với các nguồn đáng tin cậy.

**Tính toàn vẹn (Integrity):**

* Lý do: Duy trì tính toàn vẹn của các quan hệ giữa các bảng dữ liệu (ví dụ: giữa bảng Orders và Order Items, giữa bảng Customers và Orders) để tránh lỗi và mâu thuẫn.
* Cách thực hiện: Sử dụng các khóa chính và khóa ngoại trong SQL Server, và các ràng buộc toàn vẹn để đảm bảo dữ liệu không bị trùng lặp hoặc bị hỏng. Ví dụ, đảm bảo mỗi đơn hàng ("order\_id") chỉ liên kết với một khách hàng ("customer\_id") duy nhất.

**Tính đầy đủ (Completeness):**

* Lý do: Đảm bảo rằng tất cả các trường dữ liệu cần thiết đều có giá trị, giúp phân tích dữ liệu toàn diện và chính xác. Ví dụ, các trường "order\_approved\_at" trong bảng Orders, "product\_category name" trong bảng Products không được phép null.
* Cách thực hiện: Xử lý các giá trị NULL và các trường bị bỏ trống bằng cách sử dụng các kỹ thuật như điền giá trị trung bình, giá trị phổ biến nhất, hoặc sử dụng các thuật toán dự đoán.

**Tính nhất quán (Consistency):**

* Lý do: Dữ liệu nhất quán giúp đảm bảo rằng các phân tích và báo cáo dựa trên dữ liệu là đáng tin cậy.
* Cách thực hiện: Định dạng các trường dữ liệu một cách nhất quán trong toàn bộ cơ sở dữ liệu. Ví dụ, định dạng ngày tháng phải nhất quán (ví dụ: YYYY-MM-DD HH:MM:SS), và các giá trị tiền tệ phải sử dụng cùng một đơn vị (ví dụ: BRL).

**Tính kịp thời (Timeliness):**

* Lý do: Dữ liệu kịp thời giúp đưa ra các quyết định dựa trên thông tin mới nhất và chính xác, đặc biệt là các thông tin về trạng thái đơn hàng ("order\_status").
* Cách thực hiện: Thiết lập các quy trình cập nhật dữ liệu định kỳ (ví dụ: hàng ngày) hoặc theo thời gian thực nếu có thể.

**Tính dễ hiểu (Understandability):**

* Lý do: Dữ liệu phải dễ hiểu và dễ sử dụng để các bên liên quan có thể sử dụng dữ liệu một cách hiệu quả.
* Cách thực hiện: Sử dụng các tên cột và các giá trị dữ liệu rõ ràng, dễ hiểu và không gây nhầm lẫn. Ví dụ, sử dụng tên cột đầy đủ và có ý nghĩa, sử dụng các giá trị phân loại rõ ràng và nhất quán.

**Tính truy xuất (Accessibility):**

* Lý do: Dữ liệu phải dễ dàng truy cập và sử dụng bởi các bên liên quan, bao gồm cả việc đảm bảo dữ liệu có thể được truy cập một cách an toàn và bảo mật.
* Cách thực hiện: Cung cấp quyền truy cập dữ liệu cho các bên liên quan thông qua các công cụ phù hợp như Power BI, SQL Server Management Studio, đồng thời thiết lập các cơ chế bảo mật để kiểm soát truy cập và ngăn chặn truy cập trái phép.

### các bước làm sạch dữ liệu

#### Trình bày các bước làm sạch

Làm sạch dữ liệu là một bước quan trọng trong quá trình phân tích dữ liệu, giúp đảm bảo tính chính xác và đáng tin cậy của dữ liệu. Các bước làm sạch dữ liệu thường bao gồm:

* Xác định và xóa bỏ các giá trị thiếu: Kiểm tra và xử lý các giá trị bị thiếu bằng cách xóa các hàng chứa giá trị thiếu hoặc thay thế bằng giá trị trung bình, giá trị phổ biến, hoặc các kỹ thuật khác.
* Loại bỏ các giá trị ngoại lai: Xác định các giá trị ngoại lai hoặc bất thường có thể ảnh hưởng đến kết quả phân tích và loại bỏ hoặc điều chỉnh chúng.
* Chuẩn hóa dữ liệu: Đảm bảo tất cả các dữ liệu đều tuân theo một định dạng thống nhất, chẳng hạn như chuẩn hóa định dạng ngày tháng, chuyển đổi chữ hoa thành chữ thường, và loại bỏ các ký tự không cần thiết.
* Kiểm tra tính hợp lệ của dữ liệu: Đảm bảo dữ liệu tuân thủ các quy tắc và điều kiện nhất định, ví dụ như kiểm tra mã số định danh có đúng định dạng hay không, kiểm tra giá trị trong một phạm vi hợp lý.
* Xử lý dữ liệu trùng lặp: Xác định và loại bỏ các hàng dữ liệu trùng lặp để tránh tình trạng lặp lại thông tin.
* Chuyển đổi dữ liệu: Thực hiện các thao tác chuyển đổi dữ liệu như mã hóa lại các giá trị phân loại thành số, tính toán các biến mới từ các biến hiện có.

Việc làm sạch dữ liệu giúp đảm bảo rằng dữ liệu sử dụng trong phân tích là chính xác, nhất quán và sẵn sàng cho các bước xử lý tiếp theo, từ đó cải thiện độ tin cậy và tính chính xác của các kết quả phân tích.

#### Trình bày các bước làm sạch trong phạm vi dự án

**QUÁ TRÌNH LÀM SẠCH TỪNG BẢNG**

**Dữ liệu trước và sau làm sạch**

| Trước làm sạch | Sau làm sạch |
| --- | --- |
| Ngày cập nhật khi thu thập sẽ có 2 định dạng, date và general (date bên trái, text bên phải) |  |
| Lương chia làm số và cạnh tranh |  |
| Như ko có thông tin về kinh nghiệm nên thông tin ngày đăng bị lấy sang |  |
| Loại bỏ trùng lặp, bản ghi trống do ko lấy đc thông tin , loại bỏ dấu cách |  |
| Lấy thông tin cần thiết (Thành phố) |  |
| Chuyển chưa có kinh nghiệm về 0 |  |
| khoảng kinh nghiệm lấy trung bình |  |
| Chuyển not found về 0 |  |
| Xóa những hàng không có dữ liệu |  |

*Hình 3.1: Dữ liệu trước và sau làm sạch*

## chuyển đổi dữ liệu

### các kỹ thuật chuyển đổi

Kỹ thuật chuyển đổi kiểu dữ liệu (Data Type Conversion) là một quy trình quan trọng trong quá trình xử lý và chuẩn bị dữ liệu để đảm bảo tính nhất quán và chính xác của dữ liệu. Trong phân tích dữ liệu và khoa học dữ liệu, dữ liệu thường đến từ nhiều nguồn khác nhau với các định dạng và kiểu dữ liệu khác nhau, đòi hỏi việc chuyển đổi để có thể xử lý và phân tích một cách hiệu quả.

Chuyển đổi kiểu dữ liệu bao gồm việc chuyển đổi giá trị của một biến từ kiểu dữ liệu này sang kiểu dữ liệu khác. Ví dụ, dữ liệu về giá cả có thể được lưu trữ dưới dạng chuỗi (string) nhưng cần chuyển đổi sang số thập phân (float) để thực hiện các phép tính toán học. Tương tự, dữ liệu ngày tháng có thể được lưu dưới dạng chuỗi và cần chuyển đổi sang kiểu ngày tháng (datetime) để phân tích theo thời gian.

Kỹ thuật chuyển đổi kiểu dữ liệu không chỉ giúp đảm bảo tính nhất quán của dữ liệu mà còn giúp tăng cường khả năng phân tích và trực quan hóa dữ liệu, từ đó cung cấp những thông tin chính xác và có giá trị cho việc ra quyết định.

Trong phạm vi dự án này, các kỹ thuật chuyển đổi sau sẽ được áp dụng:

**Chuyển Đổi Kiểu Dữ Liệu**

* Chuyển đổi ngày tháng từ chuỗi ký tự sang kiểu datetime.
* Chuyển đổi các giá trị số từ chuỗi ký tự sang kiểu số nguyên hoặc số thực.

**Chuyển Đổi Định Dạng Dữ Liệu**

* Định dạng ngày tháng để đảm bảo thống nhất.
* Định dạng số tiền để đảm bảo thống nhất.

**Chuyển Đổi Giá Trị Phân Loại**

* Chuyển đổi tất cả các giá trị phân loại về chữ thường hoặc chữ hoa.
* Thay thế các giá trị phân loại không nhất quán hoặc không chính xác.

**Tính Toán và Chuyển Đổi Các Chỉ Số**

* Tính toán các chỉ số mới từ các cột hiện có.
* Chuyển đổi các chỉ số sang các đơn vị đo lường khác nhau nếu cần thiết.

**Xử lý dữ liệu bị thiếu**

* Xử lý các giá trị bị thiếu trong dữ liệu bằng cách thay thế bằng giá trị trung bình, giá trị phổ biến nhất, hoặc sử dụng các phương pháp nội suy.

# xử lý dữ liệu

## chuẩn hóa dữ liệu

### trình bày các bước chuẩn hóa trong dự án

Trong dự án này, quá trình chuẩn hóa dữ liệu được thực hiện trực tiếp trong Power BI để đảm bảo tính nhất quán, chính xác và sẵn sàng cho việc phân tích dữ liệu tuyển dụng tại Việt Nam. Dữ liệu được nhập vào Power BI từ các nguồn như Excel, CSV, hoặc Google Sheets, sau đó được xử lý trong Power Query Editor. Đầu tiên, các giá trị thiếu trong các cột quan trọng như Ngành nghề và Thành phố được xác định và thay thế bằng giá trị mặc định hoặc giá trị trung bình của ngành nghề tương ứng. Tiếp theo, các giá trị không hợp lệ trong cột Kinh nghiệm và Cấp bậc được kiểm tra và sửa lỗi để đảm bảo tính hợp lý. Các giá trị ngoại lai trong cột Lương max và Lương min (quá thấp hoặc quá cao) được loại bỏ để đảm bảo dữ liệu phản ánh đúng thực tế. Các cột ngày tháng như Ngày cập nhật được chuyển sang định dạng chuẩn (YYYY-MM-DD), trong khi các cột số như Lương max, Lương min được chuyển sang kiểu dữ liệu số. Dữ liệu phân loại, chẳng hạn như Ngôn ngữ và Phúc lợi, được mã hóa thành các giá trị nhị phân hoặc chia nhỏ thành các danh mục riêng lẻ để dễ dàng phân tích. Ngoài ra, các trường thông tin mới được tạo ra như Mức lương trung bình (tính trung bình giữa Lương max và Lương min) và Thời gian đăng tuyển (khoảng cách từ Ngày cập nhật đến ngày hiện tại). Để đảm bảo tính nhất quán, các ký tự thừa, khoảng trắng và ký tự đặc biệt trong các cột văn bản được loại bỏ, đồng thời giá trị văn bản được chuẩn hóa về chữ thường hoặc chữ hoa. Sau khi hoàn tất, dữ liệu được lưu và sử dụng trong Power BI để trực quan hóa và phân tích, giúp cung cấp thông tin rõ ràng, hỗ trợ việc ra quyết định một cách hiệu quả.

## mô hình hóa dữ liệu

### các loại mô hình hóa

Mô hình hóa dữ liệu là quá trình tạo ra các biểu đồ và cấu trúc dữ liệu để thể hiện mối quan hệ giữa các thành phần dữ liệu trong hệ thống. Có nhiều loại mô hình hóa dữ liệu, mỗi loại phục vụ các mục đích khác nhau trong việc thiết kế và quản lý cơ sở dữ liệu. Các loại mô hình hóa dữ liệu chính bao gồm mô hình khái niệm, mô hình logic và mô hình vật lý.

* **Mô hình khái niệm (Conceptual Data Model):** Đây là mô hình cấp cao nhất, tập trung vào việc định nghĩa các thực thể và mối quan hệ giữa chúng mà không quan tâm đến chi tiết kỹ thuật. Nó thường được sử dụng trong giai đoạn đầu của dự án để phác thảo cấu trúc dữ liệu và xác định các yêu cầu nghiệp vụ.
* **Mô hình logic (Logical Data Model):** Mô hình này chi tiết hơn mô hình khái niệm và tập trung vào cách các thực thể và mối quan hệ được biểu diễn dưới dạng bảng và cột mà không phụ thuộc vào hệ quản trị cơ sở dữ liệu cụ thể. Nó bao gồm các ràng buộc, khóa chính và khóa ngoại để đảm bảo tính toàn vẹn dữ liệu.
* **Mô hình vật lý (Physical Data Model):** Đây là mô hình chi tiết nhất, bao gồm cấu trúc lưu trữ thực tế của dữ liệu trong hệ quản trị cơ sở dữ liệu cụ thể. Mô hình này định nghĩa chi tiết về cách dữ liệu được lưu trữ, chỉ mục, phân vùng, và các yếu tố kỹ thuật khác.

Trong dự án phân tích dữ liệu, cả mô hình logic và mô hình vật lý đều đóng vai trò quan trọng trong việc thiết kế và triển khai cơ sở dữ liệu. Mô hình logic tập trung vào việc biểu diễn các thực thể kinh doanh và mối quan hệ giữa chúng mà không phụ thuộc vào công cụ hay hệ quản trị cơ sở dữ liệu cụ thể. Ví dụ, mô hình logic sẽ xác định các thực thể như Tên công việc (Tên vị trí tuyển dụng), Tên công ty (Tên doanh nghiệp hoặc tổ chức tuyển dụng), Ngành nghề (Lĩnh vực hoạt động hoặc ngành mà công việc thuộc về), cùng với các thông tin như Ngày cập nhật (Ngày gần nhất thông tin công việc được cập nhật), và mối quan hệ giữa chúng.

Trong khi đó, mô hình vật lý được xây dựng dựa trên mô hình logic và tập trung vào việc triển khai thực tế. Ví dụ, mô hình vật lý xác định kiểu dữ liệu cho các cột như Lương max, Lương min, và Lương cạnh tranh (Mức lương đề xuất hoặc thỏa thuận, có thể là một khoảng hoặc cụ thể), đảm bảo các giá trị này được định nghĩa đúng kiểu dữ liệu số để hỗ trợ tính toán. Các thông tin như Kinh nghiệm (Số năm kinh nghiệm yêu cầu), Cấp bậc (Vị trí trong hệ thống cấp bậc của công ty), và Phúc lợi (Các quyền lợi dành cho nhân viên) cũng được mã hóa phù hợp để phục vụ việc phân tích. Ngoài ra, mô hình vật lý còn xử lý các thông tin khác như Mô tả công việc, Yêu cầu công việc, Trình độ học vấn (Trình độ giáo dục tối thiểu mà công việc yêu cầu), và Ngôn ngữ (Ngôn ngữ sử dụng trong công việc), tối ưu hóa hiệu suất truy vấn và đảm bảo tính nhất quán của dữ liệu.

Sử dụng kết hợp mô hình logic và vật lý mang lại nhiều lợi ích cho dự án, bao gồm:

* **Chuẩn hóa dữ liệu:** Giúp giảm thiểu dữ liệu trùng lặp và đảm bảo tính nhất quán.
* **Tối ưu hóa hiệu suất:** Cải thiện tốc độ truy vấn và truy xuất dữ liệu.
* **Dễ dàng bảo trì:** Đơn giản hóa việc bảo trì và cập nhật cơ sở dữ liệu.
* **Tăng khả năng mở rộng:** Giúp cơ sở dữ liệu dễ dàng mở rộng khi dữ liệu tăng lên.

Tóm lại, mô hình logic và mô hình vật lý là những bước thiết yếu trong việc xây dựng một hệ thống cơ sở dữ liệu hiệu quả cho Olist, đảm bảo dữ liệu được tổ chức một cách logic và tối ưu hóa cho việc phân tích và ra quyết định kinh doanh.

### các tiêu chí đánh giá mô hình dữ liệu

**Các tiêu chí đánh giá mô hình dữ liệu tốt**

Một mô hình dữ liệu tốt cần đáp ứng nhiều tiêu chí khác nhau để đảm bảo tính hiệu quả, toàn vẹn, và khả năng mở rộng. Dưới đây là các tiêu chí đánh giá một mô hình dữ liệu tốt:

* **Tính chính xác (Accuracy):** Mô hình dữ liệu cần phản ánh đúng thực tế và đáp ứng các yêu cầu nghiệp vụ của hệ thống tuyển dụng. Các bảng dữ liệu như Tên công việc (Tên vị trí tuyển dụng), Tên công ty (Tên doanh nghiệp hoặc tổ chức tuyển dụng), và Ngành nghề (Lĩnh vực hoạt động hoặc ngành mà công việc thuộc về) phải được thiết kế chính xác để đảm bảo tính toàn diện của thông tin. Mỗi bản ghi trong hệ thống, ví dụ, một công việc cụ thể, phải đảm bảo tính chính xác về các trường như Ngày cập nhật, Lương max, Lương min, và Lương cạnh tranh, phản ánh đúng các giá trị mà nhà tuyển dụng đề xuất.
* **Tính toàn vẹn (Integrity):** Để đảm bảo không xảy ra lỗi hoặc sự thiếu nhất quán, mô hình cần sử dụng các ràng buộc toàn vẹn như khóa chính và khóa ngoại. Ví dụ, trong hệ thống, mỗi vị trí tuyển dụng phải có mã công việc duy nhất và thông tin liên quan như Cấp bậc, Phúc lợi, và Thành phố phải liên kết chính xác giữa các bảng dữ liệu.
* **Tính nhất quán (Consistency):** Dữ liệu trong hệ thống cần đảm bảo không có sự mâu thuẫn giữa các bảng hoặc các trường. Ví dụ, định dạng ngày tháng trong cột Ngày cập nhật cần thống nhất ở định dạng chuẩn (YYYY-MM-DD). Các trường như Lương max, Lương min, và các thông tin tiền lương khác cũng phải sử dụng chung một đơn vị tiền tệ để duy trì tính nhất quán trong toàn hệ thống.
* **Tính đầy đủ (Completeness):** Mô hình dữ liệu phải bao gồm tất cả các thực thể và mối quan hệ cần thiết để đáp ứng yêu cầu nghiệp vụ. Các trường như Yêu cầu công việc, Nhóm công việc, Ngôn ngữ, và Trình độ học vấn cần được lưu trữ đầy đủ để đảm bảo hệ thống cung cấp đủ thông tin cho cả nhà tuyển dụng và ứng viên.
* **Tính mở rộng (Scalability):** Hệ thống cần được thiết kế để dễ dàng mở rộng khi có nhu cầu. Ví dụ, các bảng có thể dễ dàng thêm các trường mới như Địa điểm làm việc chi tiết hoặc mở rộng các tính năng, chẳng hạn thêm danh sách các công việc tương tự, mà không ảnh hưởng đến cấu trúc dữ liệu hiện tại.
* **Tính hiệu quả (Efficiency):** Mô hình phải hỗ trợ xử lý và truy vấn dữ liệu hiệu quả, giảm thiểu thời gian phản hồi hệ thống. Ví dụ, việc tạo chỉ mục trên các trường như Tên công ty, Ngành nghề, hoặc Kinh nghiệm sẽ giúp tăng tốc độ truy vấn khi tìm kiếm các công việc phù hợp.

**Tính dễ hiểu (Understandability)**

* Mô tả: Mô hình phải dễ hiểu và dễ sử dụng cho các nhà phát triển, nhà quản lý dữ liệu và người dùng cuối.
* Ví dụ: Các tên bảng và cột phải rõ ràng và dễ hiểu.

**Mô hình đang dùng trong dự án đáp ứng các tiêu chí nào?**

Mô hình dữ liệu hiện tại trong dự án đã đáp ứng được hầu hết các tiêu chí quan trọng như tính chính xác, tính toàn vẹn, tính nhất quán, tính đầy đủ, tính mở rộng và tính hiệu quả. Điều này đảm bảo rằng cơ sở dữ liệu không chỉ chính xác và nhất quán mà còn có khả năng mở rộng và hiệu quả trong việc xử lý dữ liệu, hỗ trợ tốt cho các yêu cầu phân tích và báo cáo của dự án.

**Tính chính xác (Accuracy):** Mô hình phản ánh chính xác các thực thể và mối quan hệ trong dữ liệu.

**Tính toàn vẹn (Integrity):** Mô hình sử dụng các khóa chính và khóa ngoại để đảm bảo tính toàn vẹn dữ liệu. Các ràng buộc toàn vẹn đảm bảo rằng không có dữ liệu trùng lặp hoặc mất mát trong quá trình nhập và xử lý dữ liệu.

**Tính nhất quán (Consistency):** Định dạng và kiểu dữ liệu được chuẩn hóa trong toàn bộ cơ sở dữ liệu. Ví dụ, các cột ngày tháng và số liệu được định dạng nhất quán, đảm bảo tính nhất quán dữ liệu.

**Tính đầy đủ (Completeness):** Mô hình bao gồm tất cả các thông tin cần thiết.

**Tính mở rộng (Scalability):** Cấu trúc bảng và mối quan hệ giữa chúng cho phép mở rộng dễ dàng. Ví dụ, có thể thêm các bảng mới hoặc mở rộng các bảng hiện có mà không làm gián đoạn hệ thống.

**Tính hiệu quả (Efficiency):** Mô hình được thiết kế để hỗ trợ truy vấn và xử lý dữ liệu hiệu quả. Các chỉ mục và thiết kế bảng tối ưu giúp giảm thiểu thời gian truy xuất và sử dụng tài nguyên hệ thống.

### trình bày các bước mô hình hóa

Quá trình phân tích và xác định các bảng cần thiết giúp chuẩn hóa cơ sở dữ liệu, loại bỏ dữ liệu dư thừa, tăng tính toàn vẹn và hiệu quả trong quản lý dữ liệu. Việc chia nhỏ dữ liệu thành các bảng và thiết lập mối quan hệ giữa chúng giúp cơ sở dữ liệu dễ bảo trì và mở rộng hơn trong tương lai. Triển khai mô hình vật lý cụ thể với các lệnh SQL giúp đảm bảo rằng cơ sở dữ liệu được cấu trúc một cách chính xác và sẵn sàng cho việc sử dụng trong các ứng dụng thực tế.

**Bước 1: Xác định các thực thể và thuộc tính**

**Công việc (Jobs):**

* Thuộc tính chính: Mã công việc (job\_id).
* Thuộc tính bổ sung: Tên công việc, Ngày cập nhật, Lương Max, Lương Min, Lương cạnh tranh, Kinh nghiệm, Mã nhóm công việc.

**Công ty (Companies):**

* Thuộc tính chính: Mã công ty (company\_id).
* Thuộc tính bổ sung: Tên công ty.

**Nhóm công việc (Job Groups):**

* Thuộc tính chính: Mã nhóm công việc (job\_group\_id).
* Thuộc tính bổ sung: Tên nhóm công việc.

**Bằng cấp (Qualifications):**

* Thuộc tính chính: Mã bằng cấp (qualification\_id).
* Thuộc tính bổ sung: Loại bằng cấp.

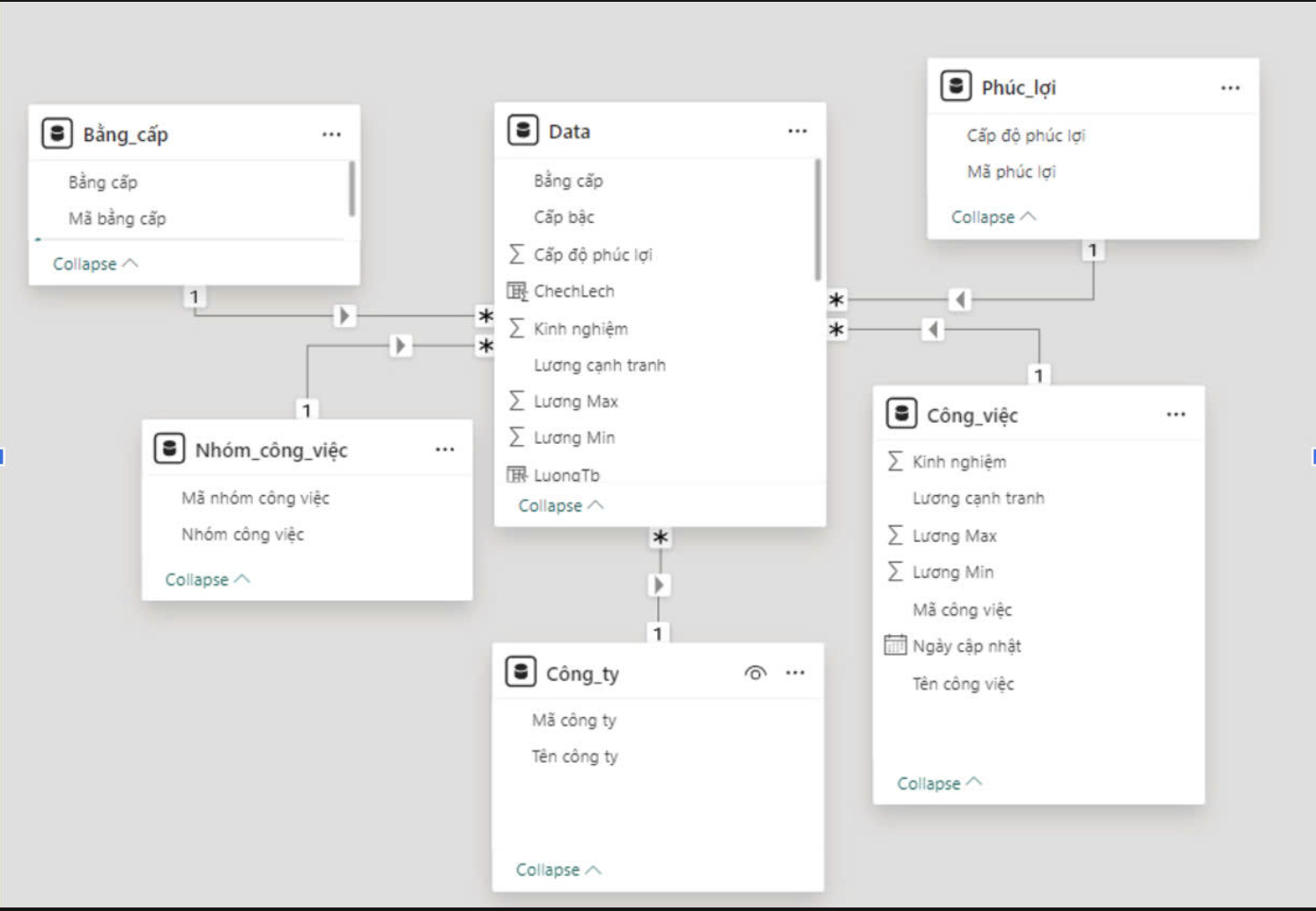
**Phúc lợi (Benefits):**

* Thuộc tính chính: Mã phúc lợi (benefit\_id).
* Thuộc tính bổ sung: Cấp độ phúc lợi.

**Data:**

* Thuộc tính chính: Không có (liên kết thông qua các khóa ngoại).
* Thuộc tính bổ sung: Cấp bậc, Cấp độ phúc lợi, Kinh nghiệm, Lương Min, Lương Max, Lương cạnh tranh, Chênh lệch lương, Mã công việc, Mã công ty, Mã nhóm công việc, Mã bằng cấp.

**Bước 2: Xác định các mối quan hệ**

****

*Hình 4.1: Mối quan hệ*

**Bảng Bằng\_cấp → Data:**

* Mối quan hệ: 1:N (Một Bằng cấp có thể liên kết với nhiều bản ghi trong bảng Data).
* Khóa: Cột Bằng cấp trong cả hai bảng.

**Bảng Nhóm Công Việc → Data:**

* Mối quan hệ: 1:N (Một Nhóm công việc có thể liên kết với nhiều bản ghi trong bảng Data).
* Khóa: Cột Mã nhóm công việc trong cả hai bảng.

**Bảng Phúc Lợi → Data:**

* Mối quan hệ: 1:N (Một Phúc lợi có thể liên kết với nhiều bản ghi trong bảng Data).
* Khóa: Cột Mã phúc lợi trong cả hai bảng.

**Bảng Công Việc → Data:**

* Mối quan hệ: 1:N (Một Công việc có thể liên kết với nhiều bản ghi trong bảng Data).
* Khóa: Cột Mã công việc trong cả hai bảng.

**Bảng Công Ty → Data:**

* Mối quan hệ: 1:N (Một Công ty có thể liên kết với nhiều bản ghi trong bảng Data).
* Khóa: Cột Mã công ty trong cả hai bảng.

**Bước 3: Kiểm tra và tối ưu hóa thiết kế**

Kiểm tra toàn vẹn dữ liệu: Đảm bảo rằng thiết kế bảng và các mối quan hệ của bạn đảm bảo tính toàn vẹn và chính xác của dữ liệu.

Tối ưu hóa: Đảm bảo rằng cơ sở dữ liệu của bạn có thể xử lý khối lượng dữ liệu lớn một cách hiệu quả và các truy vấn sẽ được thực thi nhanh chóng.

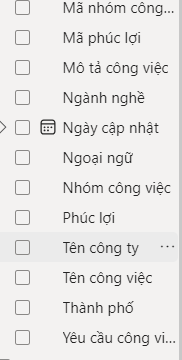
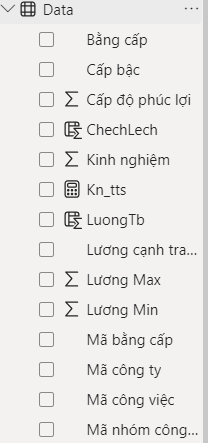
**Bước 4: Đảm bảo tính mở rộng và bảo trì**

Thiết kế cơ sở dữ liệu cần đảm bảo tính linh hoạt và dễ dàng mở rộng trong tương lai. Các thay đổi trong yêu cầu kinh doanh có thể dẫn đến việc điều chỉnh cấu trúc cơ sở dữ liệu.

## xử lý dữ liệu dax

### measure

#### Tạo bang data



*Hình 4.2: Danh sách các measure được sử dụng trong báo cáo*



*Hình 4.3: Measure tính kinh nghiệm thực tập sinh*



*Hình 4.4: Measure tính lương trung bình*



*Hình 4.5: Công thức tính lương chênh lệch*



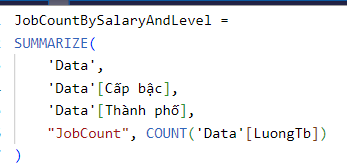
*Hình 4.6: Công thức tính lương trung bình theo thành phố*



*Hình 4.7: Measure tính lương trung bình theo cấp bậc*



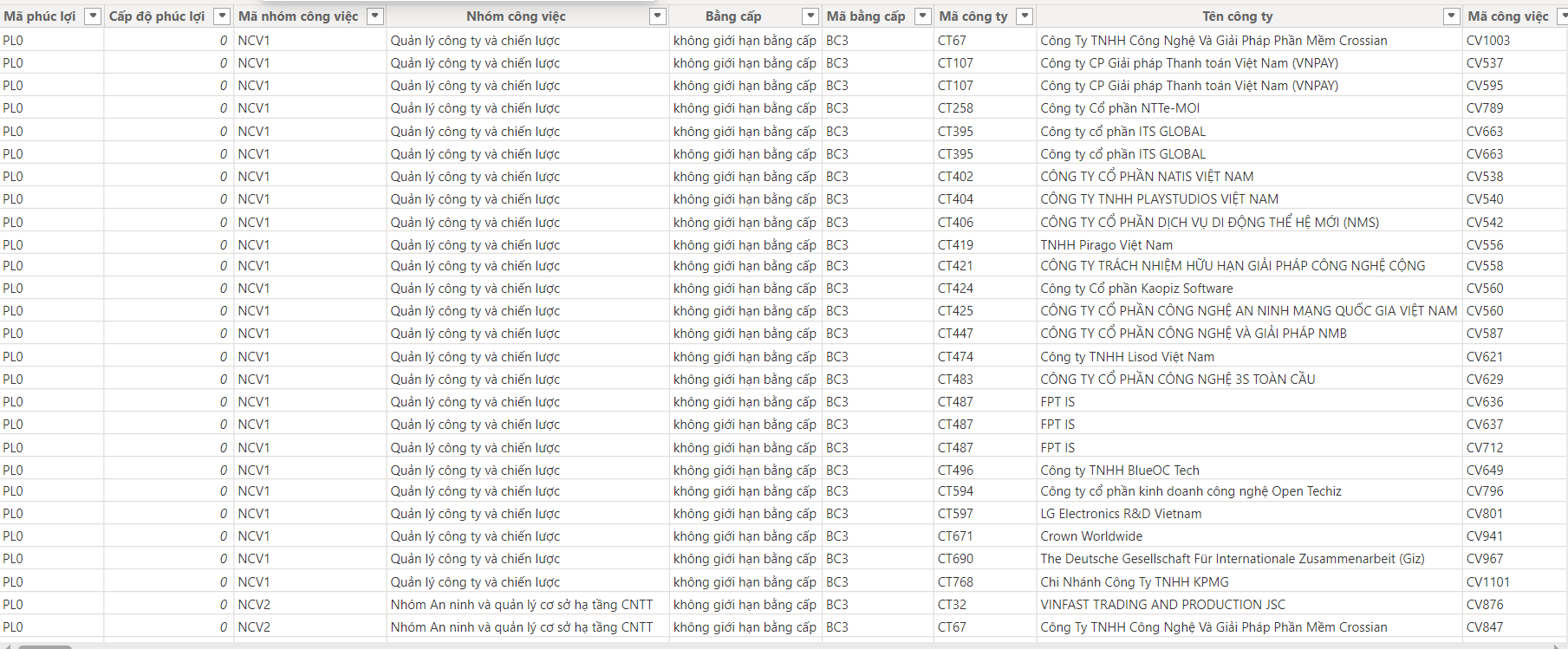
*Hình 4.8: Measure top 3 thành phố với mức lương cao nhất*



*Hình 4.9: Measure tính số lượng công việc theo cấp bậc*

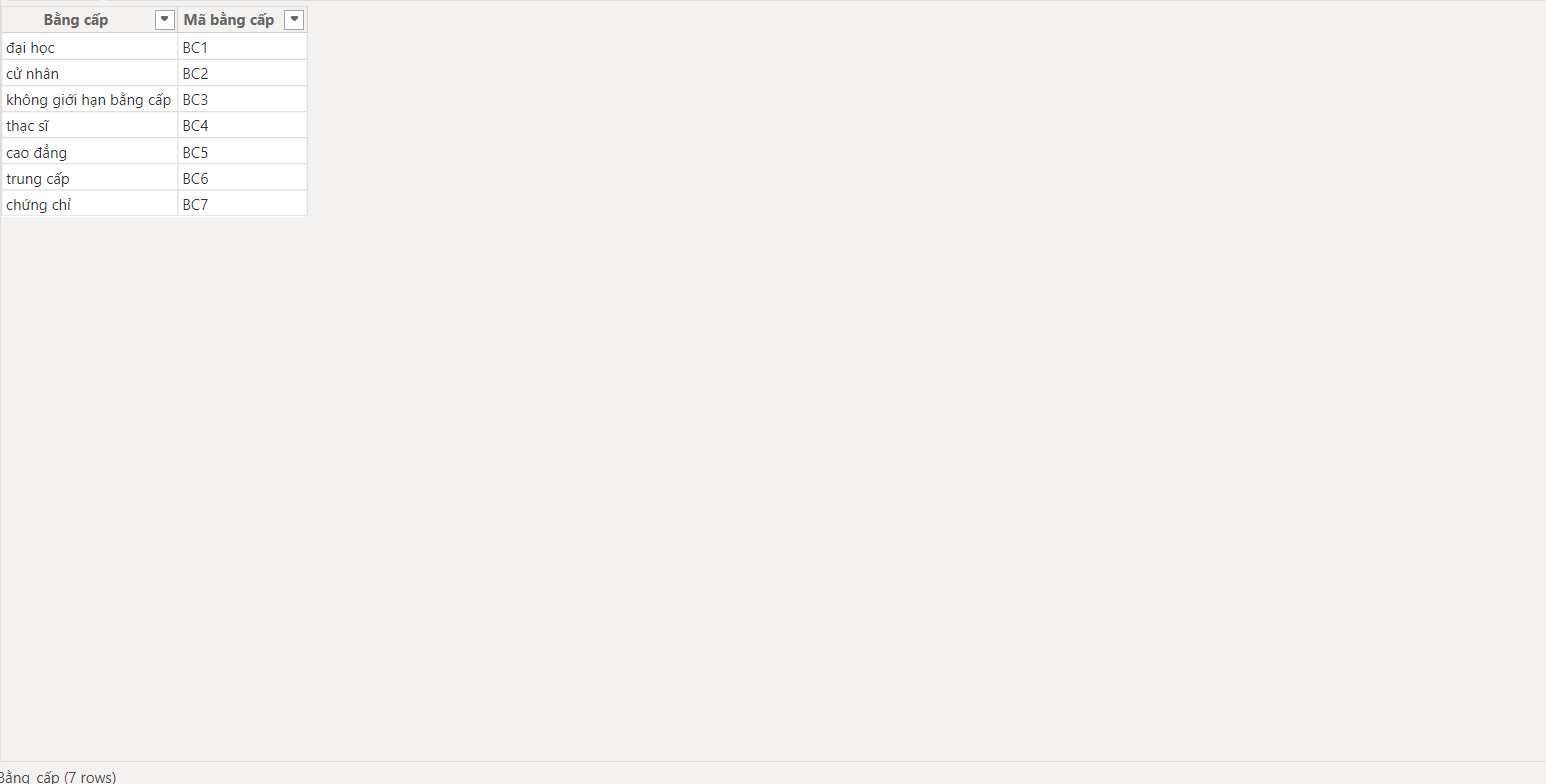
### calculated column

#### Bảng Data (Fact)



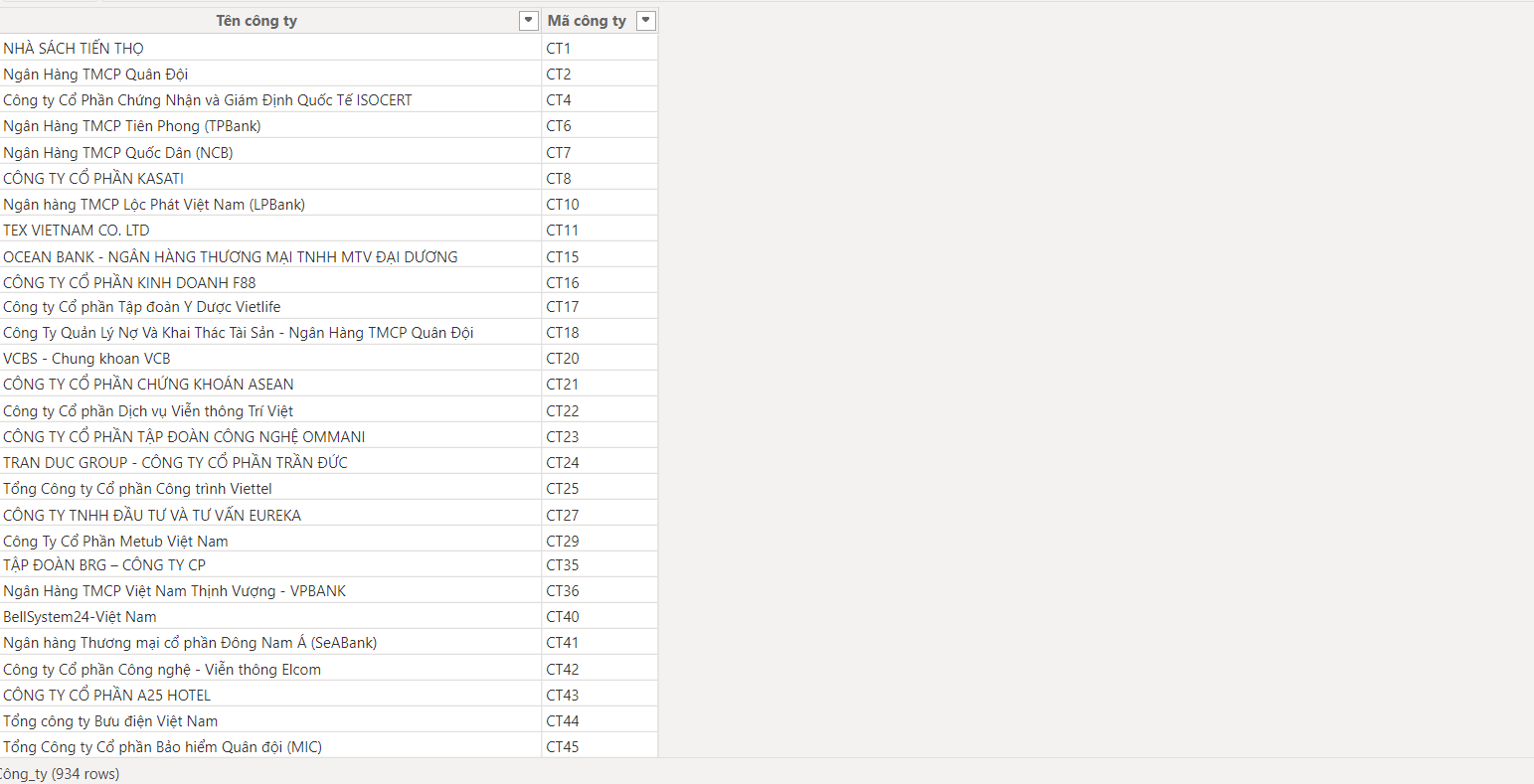
*Hình 4.10: Bảng Data*

#### Bảng bằng cấp



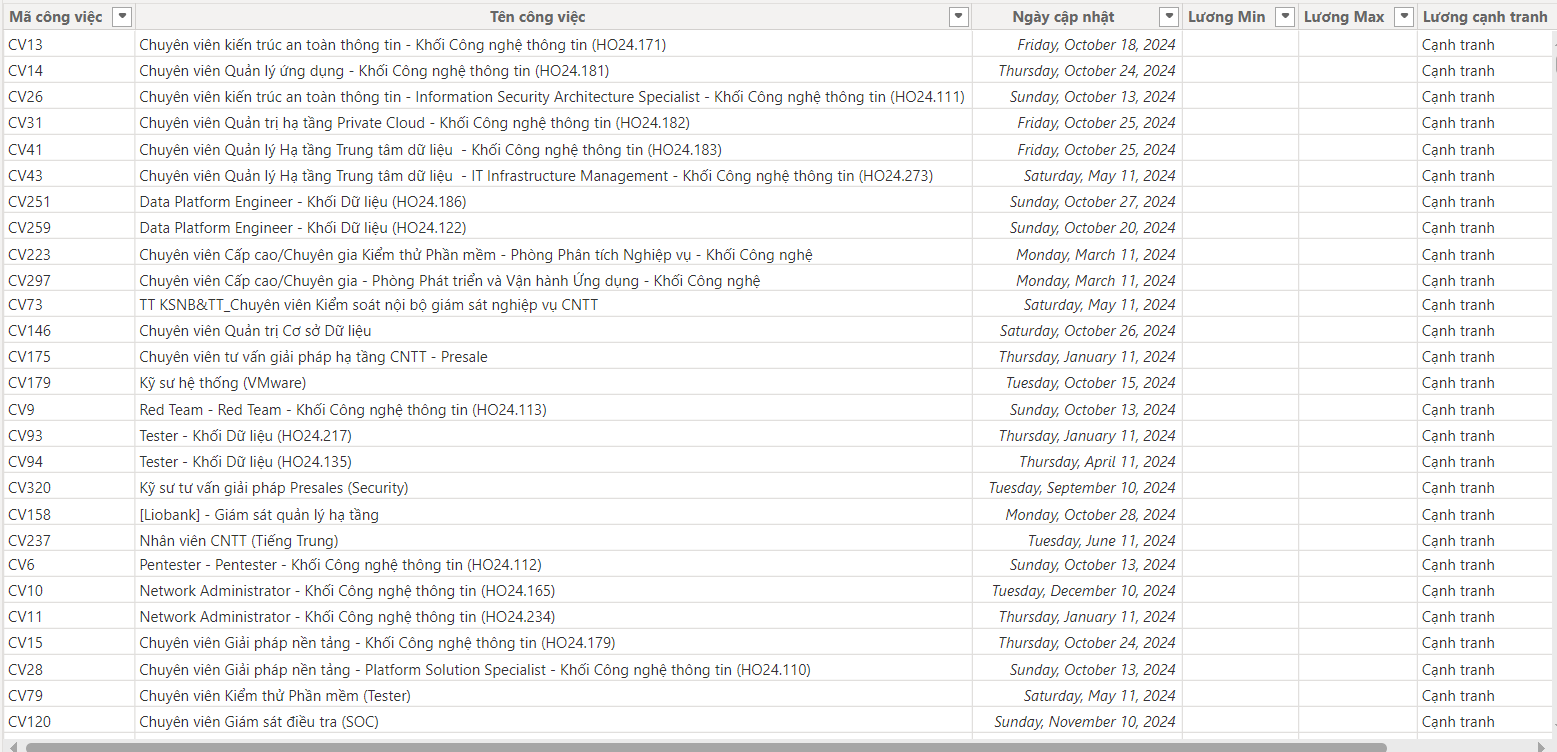
*Hình 4.11: Bảng bằng cấp*

#### Bảng công ty



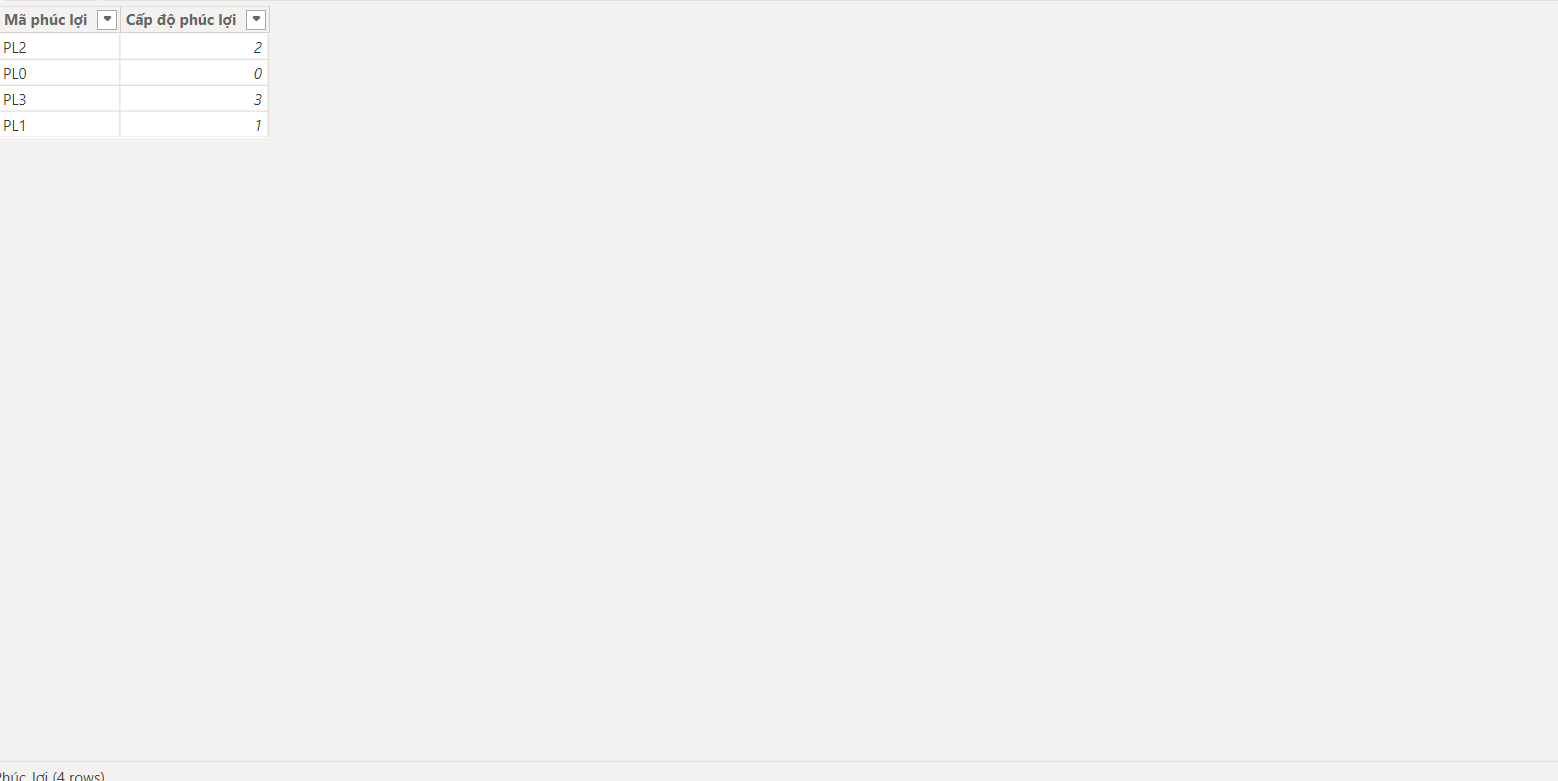
*Hình 4.12: Bảng công ty*

#### Bảng công việc



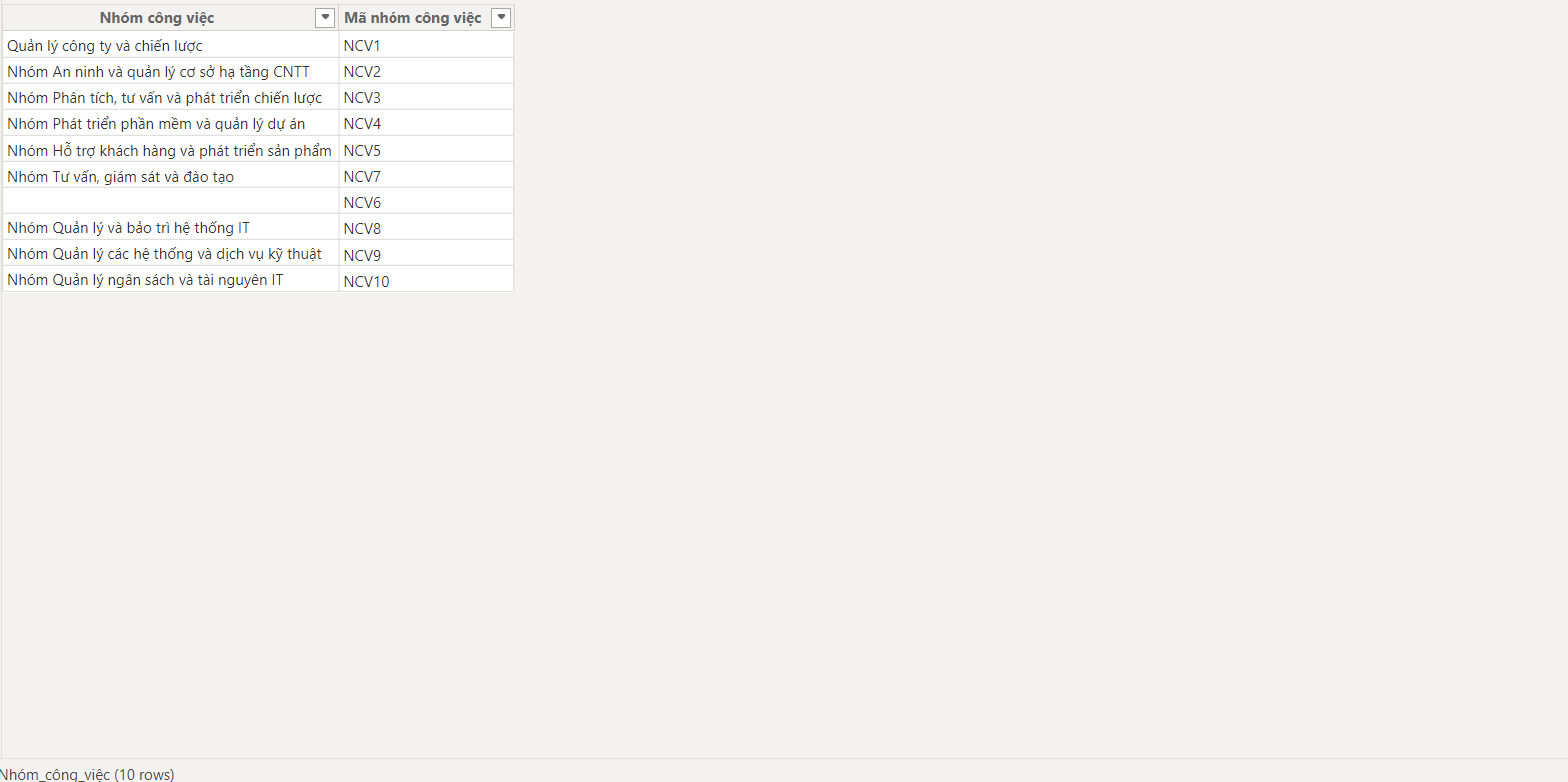
*Hình 4.13: Bảng công việc*

#### Bảng phúc lợi



*Hình 4.14: Bảng phúc lợi*

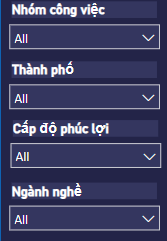
#### Bảng nhóm công việc



*Hình 4.15: Bảng nhóm công việc*

### filter

#### Tạo filter chọn nhóm công việc, thành phố, cấp độ phúc lợi, ngành nghề



*Hình 4.16: Bộ lọc chọn nhóm công việc, thành phố, cấp độ phúc lợi, ngành nghề*

#### Tạo filter chọn Ngày



*Hình 4.17: Bộ lọc chọn ngày*

# trực quan hóa dữ liệu

## các kỹ thuật trực quan hóa

Trực quan hóa dữ liệu là quá trình sử dụng các biểu đồ và đồ thị để truyền đạt thông tin một cách trực quan, dễ hiểu. Có nhiều kỹ thuật trực quan hóa dữ liệu khác nhau, mỗi kỹ thuật phù hợp với các loại dữ liệu và mục đích phân tích cụ thể. Dưới đây là một số kỹ thuật trực quan hóa phổ biến và kỹ thuật được áp dụng cho dự án này.

**Temporal data visualization (Trực quan hóa dữ liệu chuỗi thời gian)**

Trực quan hóa dữ liệu chuỗi thời gian (Temporal Data Visualization) là kỹ thuật trình bày các đối tượng dữ liệu theo chiều thời gian một cách trực quan và dễ hiểu. Kỹ thuật này thường sử dụng các loại biểu đồ như biểu đồ đường, biểu đồ cột và dòng thời gian để minh họa các thay đổi, xu hướng, và sự kiện xảy ra liên tục trong một khoảng thời gian nhất định. Ví dụ, biểu đồ đường có thể được sử dụng để biểu thị sự thay đổi của giá bán xe ô tô theo từng tháng hoặc từng năm, giúp người xem dễ dàng nhận thấy các xu hướng tăng giảm, các điểm biến động, và các chu kỳ thời gian cụ thể. Trực quan hóa dữ liệu chuỗi thời gian giúp người dùng theo dõi các biến động trong dữ liệu theo thời gian, từ đó đưa ra các quyết định dựa trên những hiểu biết sâu sắc và có căn cứ. Đây là một công cụ mạnh mẽ trong phân tích dữ liệu, đặc biệt hữu ích cho các lĩnh vực yêu cầu theo dõi xu hướng và dự báo như kinh doanh, tài chính và quản lý.

**Hierarchical data visualization (Trực quan hóa dữ liệu phân cấp)**

Trực quan hóa dữ liệu phân cấp (Hierarchical Data Visualization) là kỹ thuật dùng để trình bày các nhóm hoặc tập hợp các mục có liên kết chung với một mục cha, giúp hiển thị cấu trúc phân cấp và mối quan hệ giữa các phần tử dữ liệu. Các dạng trực quan phổ biến cho dữ liệu phân cấp bao gồm cây phân cấp, sơ đồ cây, và biểu đồ cây. Ví dụ, cây dữ liệu có thể được sử dụng để biểu thị lượng dữ liệu về hàng tồn kho, trong đó có nút cha đại diện cho danh mục lớn như "quần áo" và các nút con đại diện cho các mục nhỏ hơn như "áo sơ mi", "quần dài", và "tất". Kỹ thuật này giúp người dùng dễ dàng hiểu được cấu trúc phân cấp của dữ liệu, nhìn thấy mối quan hệ và sự phụ thuộc giữa các phần tử, từ đó hỗ trợ việc phân tích, ra quyết định và quản lý dữ liệu hiệu quả hơn. Trực quan hóa dữ liệu phân cấp đặc biệt hữu ích trong các lĩnh vực như quản lý dự án, tổ chức dữ liệu, và phân tích hệ thống phức tạp.

**Network data visualization (Trực quan hóa dữ liệu mạng)**

Trực quan hóa dữ liệu mạng (Network Data Visualization) là kỹ thuật biểu diễn dữ liệu dưới dạng các điểm và mối liên kết giữa chúng trên một đồ thị, giúp hiển thị rõ ràng các mối quan hệ và tương tác phức tạp trong mạng lưới dữ liệu. Các dạng biểu đồ phổ biến cho dữ liệu mạng bao gồm biểu đồ phân tán, biểu đồ bong bóng, và đám mây từ. Ví dụ, biểu đồ phân tán có thể hiển thị mối quan hệ giữa hai biến, trong khi biểu đồ bong bóng thêm một yếu tố dữ liệu thứ ba thông qua kích thước của bong bóng. Đám mây từ trình bày tần suất xuất hiện của các từ bằng cách sử dụng các từ có kích cỡ khác nhau, giúp nhận diện nhanh chóng các từ quan trọng. Trực quan hóa dữ liệu mạng giúp người dùng hiểu được cấu trúc và động lực của các mạng phức tạp, chẳng hạn như mạng xã hội, mạng giao thông, hoặc mối liên kết giữa các yếu tố trong một hệ thống. Kỹ thuật này đặc biệt hữu ích trong việc phân tích các mối quan hệ và tương tác, phát hiện các mẫu ẩn và xác định các yếu tố quan trọng trong mạng lưới dữ liệu.

**Multidimensional data visualization (Trực quan hóa dữ liệu đa chiều)**

Trực quan hóa dữ liệu đa chiều (Multidimensional Data Visualization) là kỹ thuật dùng để biểu diễn và phân tích dữ liệu có nhiều biến hoặc chiều, giúp người dùng dễ dàng so sánh và nhận diện các mối quan hệ giữa các yếu tố dữ liệu phức tạp. Các dạng biểu đồ phổ biến cho dữ liệu đa chiều bao gồm biểu đồ cột, biểu đồ tròn và đồ thị cột. Ví dụ, biểu đồ cột có thể so sánh các yếu tố dữ liệu khác nhau như doanh số bán hàng theo các hãng xe trong các khoảng thời gian khác nhau, biểu đồ tròn trực quan hóa tỷ lệ phần trăm của từng danh mục trong tổng thể như thị phần của các hãng xe. Kỹ thuật này cho phép người dùng theo dõi và phân tích sự thay đổi của một hoặc nhiều biến qua thời gian hoặc giữa các danh mục khác nhau, từ đó phát hiện ra các xu hướng, mẫu, và mối quan hệ quan trọng. Trực quan hóa dữ liệu đa chiều là công cụ mạnh mẽ trong các lĩnh vực như kinh doanh, tài chính, và nghiên cứu khoa học, nơi mà việc phân tích các yếu tố đa chiều là cần thiết để đưa ra các quyết định chính xác và có căn cứ.

**Geospatial data visualization (Trực quan hóa dữ liệu không gian địa lý)**

Trực quan hóa dữ liệu không gian địa lý (Geospatial Data Visualization) là kỹ thuật sử dụng các bản đồ và biểu đồ để trình bày dữ liệu liên quan đến các vị trí địa lý trong thế giới thực. Kỹ thuật này giúp biểu diễn thông tin không gian một cách trực quan, giúp người dùng dễ dàng nhận diện các mẫu, xu hướng và mối quan hệ trong dữ liệu địa lý. Các dạng trực quan phổ biến bao gồm bản đồ nhiệt, bản đồ mật độ, và bản đồ địa hình. Ví dụ, bản đồ nhiệt có thể được sử dụng để hiển thị lượng khách hàng ghé thăm các chi nhánh bán lẻ khác nhau, với màu sắc đậm nhạt biểu thị mật độ khách hàng. Bản đồ địa hình có thể minh họa các đặc điểm địa lý và các hiện tượng tự nhiên. Trực quan hóa dữ liệu không gian địa lý không chỉ giúp hiểu rõ hơn về thông tin không gian mà còn hỗ trợ việc ra quyết định dựa trên vị trí, tối ưu hóa các chiến lược kinh doanh và quản lý tài nguyên. Kỹ thuật này đặc biệt hữu ích trong các lĩnh vực như quản lý đô thị, logistics, marketing địa phương, và nghiên cứu môi trường.

**Kỹ thuật đang được áp dụng cho dự án**

Trong dự án này, chúng ta tập trung vào hai kỹ thuật chính: trực quan hóa dữ liệu chuỗi thời gian để theo dõi xu hướng và sự thay đổi theo thời gian (ví dụ: doanh thu theo tháng) và trực quan hóa dữ liệu đa chiều để khám phá mối quan hệ giữa các yếu tố khác nhau (ví dụ: doanh thu theo danh mục sản phẩm).

Ngoài ra, chúng ta có thể sử dụng bổ trợ các kỹ thuật như trực quan hóa dữ liệu phân cấp (hiển thị cấu trúc danh mục sản phẩm) và trực quan hóa dữ liệu không gian địa lý (phân tích dữ liệu theo vị trí địa lý).

Việc lựa chọn các kỹ thuật này dựa trên đặc điểm dữ liệu của Olist, vừa mang tính thời gian vừa có nhiều chiều thông tin, giúp tối ưu hóa việc phân tích và trình bày dữ liệu, từ đó hỗ trợ ra quyết định kinh doanh hiệu quả.

## các nguyên tắc trực quan hóa

**5 nguyên tắc trực quan hóa dữ liệu**

**Chọn đúng loại biểu đồ:**

* Sử dụng biểu đồ đường để hiển thị xu hướng theo thời gian, ví dụ như doanh thu theo tháng, số lượng đơn hàng theo quý.
* Sử dụng biểu đồ cột để so sánh giá trị giữa các danh mục, ví dụ như so sánh doanh thu giữa các danh mục sản phẩm, số lượng đơn hàng theo từng vùng miền.
* Sử dụng biểu đồ tròn để hiển thị tỷ lệ phần trăm, ví dụ như tỷ lệ đơn hàng thành công, tỷ lệ hủy đơn.

**Không phải tất cả dữ liệu đều quan trọng:**

* Tập trung vào những dữ liệu cốt lõi, liên quan trực tiếp đến mục tiêu phân tích.
* Loại bỏ những thông tin không cần thiết, tránh làm biểu đồ trở nên rối rắm.
* Ví dụ, khi phân tích doanh thu, chỉ nên tập trung vào các thông tin về doanh thu, số lượng đơn hàng, thời gian, thay vì hiển thị thêm các chi tiết như mã sản phẩm, thông tin chi tiết về khách hàng.

**Biểu đồ thể hiện đúng tương quan số liệu thực tế:**

* Đảm bảo biểu đồ phản ánh chính xác mối quan hệ giữa các dữ liệu.
* Ví dụ, khi sử dụng biểu đồ cột, trục tung nên bắt đầu từ 0 để tránh gây hiểu nhầm về tỷ lệ.
* Sử dụng tỷ lệ và khoảng cách phù hợp trong biểu đồ để thể hiện chính xác dữ liệu.

**Sử dụng màu sắc hợp lý:**

* Chọn màu sắc dễ phân biệt, có ý nghĩa và tránh sử dụng quá nhiều màu sắc.
* Sử dụng tương phản màu để làm nổi bật các thông tin quan trọng.
* Ví dụ, sử dụng màu xanh lá cây để thể hiện sự tăng trưởng, màu đỏ để thể hiện sự sụt giảm.

**Trình bày dữ liệu đơn giản và hiệu quả:**

* Sử dụng biểu đồ đơn giản, dễ hiểu, tránh các chi tiết thừa thãi.
* Đảm bảo các nhãn, chú thích, tiêu đề rõ ràng và dễ đọc.
* Ví dụ, sử dụng biểu đồ đường đơn giản để hiển thị xu hướng doanh thu theo thời gian, thay vì sử dụng biểu đồ kết hợp phức tạp.

**Nguyên tắc quan trọng nhất: chọn đúng loại biểu đồ**

Việc chọn đúng loại biểu đồ là yếu tố then chốt để truyền tải thông tin hiệu quả. Biểu đồ phù hợp sẽ giúp người xem dễ dàng nắm bắt được ý nghĩa của dữ liệu và đưa ra quyết định chính xác.

Ví dụ, để phân tích xu hướng doanh thu theo thời gian, biểu đồ đường là lựa chọn tốt nhất. Để so sánh doanh thu giữa các danh mục sản phẩm, biểu đồ cột sẽ phù hợp hơn.

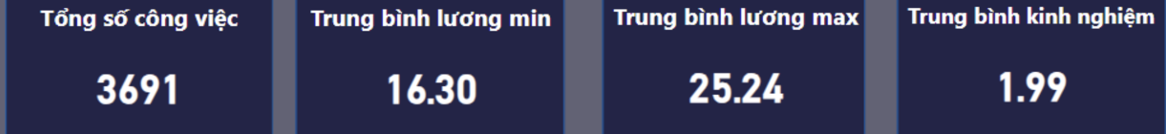
Tóm lại, việc áp dụng 5 nguyên tắc trên sẽ giúp cho việc trực quan hóa dữ liệu của Olist trở nên hiệu quả, hỗ trợ đắc lực cho quá trình phân tích và ra quyết định kinh doanh.

## trình bày tạo các report cho dự án

### tạo visual thống kê chi tiết

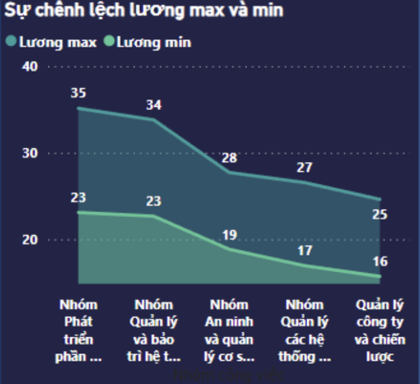
#### Trang tổng quát

* + - * 1. Tạo visual thống kê

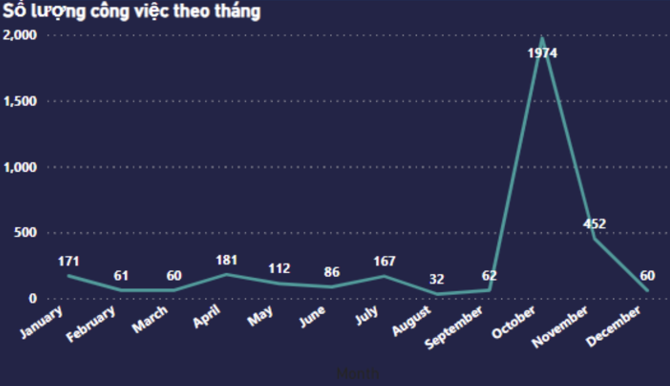


*Hình 5.1: Card tổng số công việc, trung bình lương min,trung bình lương max,trung bình kinh nghiệm*

* + - * 1. Tạo visual sự chênh lệch lương max và min

*Hình 5.2: Biểu đồ chênh lệch lương max và min*

* + - * 1. Tạo visual số lượng công việc theo tháng



*Hình 5.3: Biểu đồ số lượng công việc theo tháng*

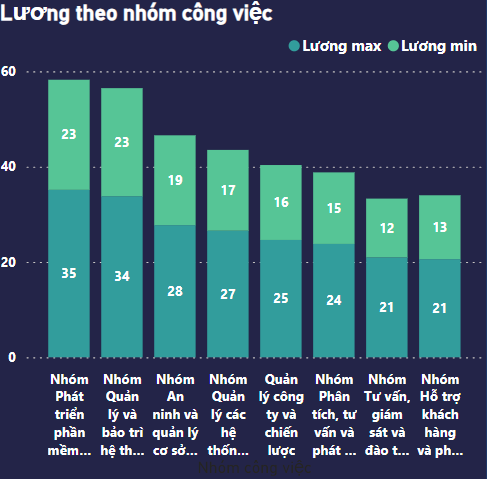
#### Trang lương

* + - * 1. Tạo visual thống kê

|  |  |
| --- | --- |
|  |  |

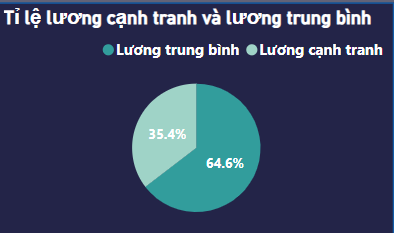
*Hình 5.4: Card thống kê về số bản tin có lương cứng , lương cạnh tranh , trung bình lương min ,trung bình lương max*

* + - * 1. Tạo visual thống kê lương max và lương min theo nhóm công việc

****

*Hình 5.5: Biểu đồ thống kê lương max và lương min theo nhóm công việc*

* + - * 1. Tạo visual so sánh tỷ lệ lương cạnh tranh và lương trung bình

****

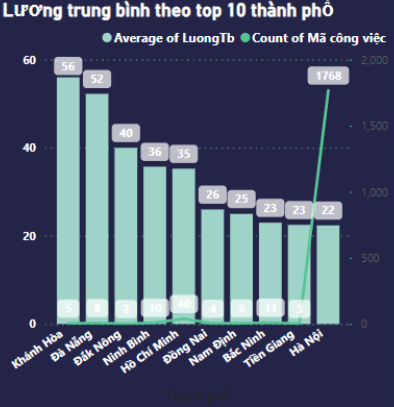
*Hình 5.6: Biểu đồ so sánh tỷ lệ lương cạnh tranh và lương trung bình*

* + - * 1. Tạo visual về lương theo số năm kinh nghiệm

****

*Hình 5.7: Biểu đồ lương theo số năm kinh nghiệm*

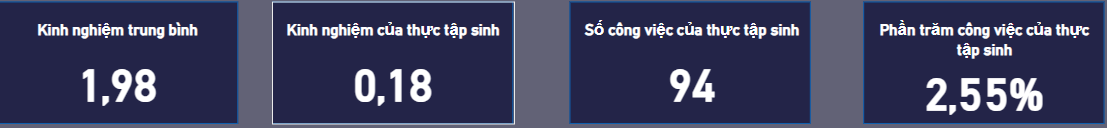
* + - * 1. Tạo visual về lương trung bình theo top 10 thành phố

****

*Hình 5.8: Biểu đồ lương trung bình theo top 10 thành phố*

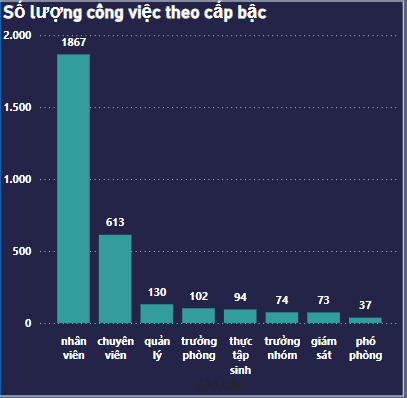
#### Trang kinh nghiệm và cấp bậc

* + - * 1. Tạo visual thống kê

****

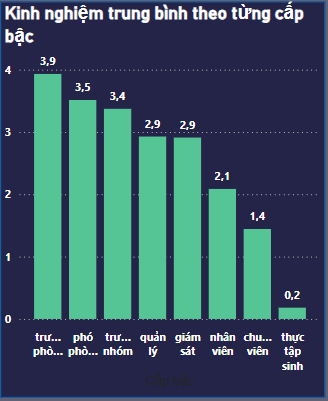
*Hình 5.9: Card thống kê về kinh nghiệm của thực tập sinh*

* + - * 1. Tạo visual phân tích số lượng công việc theo bậc

****

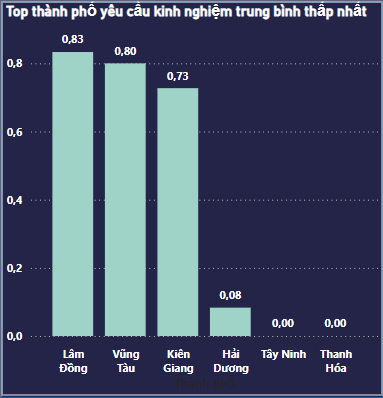
*Hình 5.10: Biểu đồ phân tích số lượng công việc theo bậc*

* + - * 1. Tạo visual phân tích kinh nghiệm trung bình theo từng cấp bậc



*Hình 5.11: Biểu đồ phân tích kinh nghiệm trung bình theo từng cấp bậc*

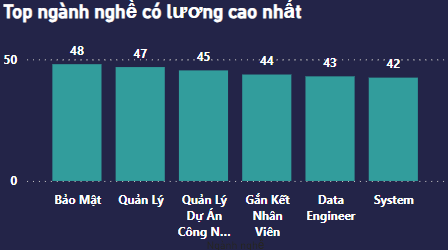
* + - * 1. Tạo visual phân tích top thành phố yêu cầu kinh nghiệm trung bình thấp nhất



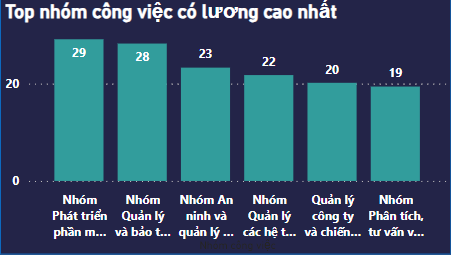
*Hình 5.12: Biểu đồ phân tích top thành phố yêu cầu kinh nghiệm trung bình thấp nhất*

#### Trang ngành nghề và nhóm công việc

* + - * 1. Tạo visual thể hiện ngành nghề có lương cao nhất

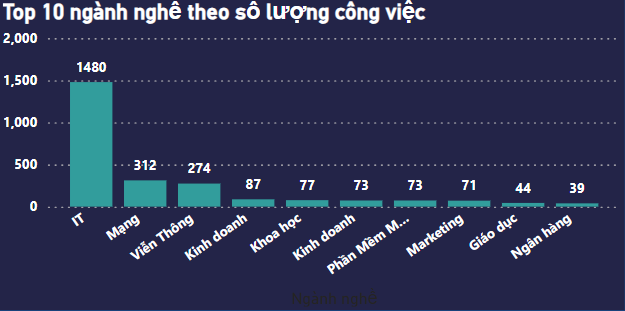


*Hình 5.13: Biểu đồ thể hiện top ngành nghề có lương cao nhất*

* + - * 1. Tạo visual thể hiện nhóm công việc có lương cao nhất

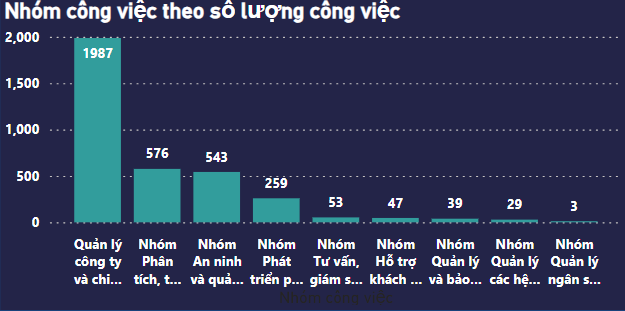
*Hình 5.14: Biểu đồ thể hiện top nhóm công việc có lương cao nhất*

* + - * 1. Tạo visual thể hiện top 10 ngành nghề theo số lượng công việc



*Hình 5.15: Biểu đồ top 10 ngành nghề theo số lượng công việc*

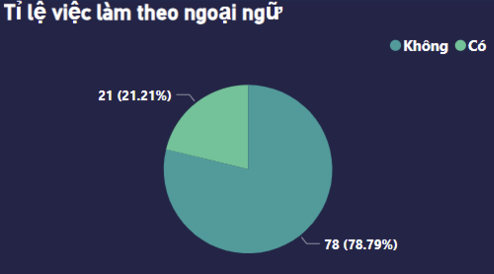
* + - * 1. Tạo visual thể hiện nhóm công việc theo số lượng công việc



*Hình 5.16: Biểu đồ thể hiện nhóm công việc theo số lượng công việc*

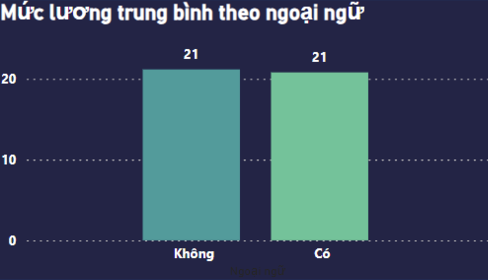
#### Trang thành phố và ngoại ngữ

* + - * 1. Tạo visual thể hiện tỉ lệ việc làm theo ngoại ngữ



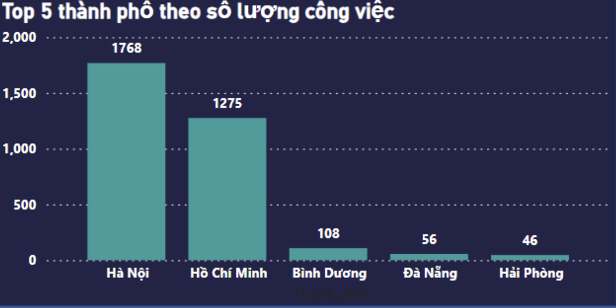
*Hình 5.17: visual thể hiện tỉ lệ việc làm theo ngoại ngữ*

* + - * 1. Tạo visual thể hiện mức lương trung bình theo ngoại ngữ



*Hình 5.18: Visual thể hiện mức lương trung bình theo ngoại ngữ*

* + - * 1. Tạo visual thể hiện top 5 thành phố theo số lượng công việc



*Hình 5.19: Biểu đồ thể hiện top 5 thành phố theo số lượng công việc*

* + - * 1. Tạo visual thể hiện top thành phố có lương cao nhất



*Hình 5.20: Biểu đồ thể hiện top thành phố có lương cao nhất*

# xây dựng báo cáo

## dashboard và report

Tối ưu hóa dashboard và report là bước quan trọng để đảm bảo thông tin được truyền tải một cách rõ ràng, dễ hiểu và hiệu quả nhất đến người sử dụng. Tối ưu hóa dashboard và report không chỉ giúp cải thiện hiệu suất mà còn đảm bảo thông tin được trình bày một cách rõ ràng, dễ hiểu và hấp dẫn. Bằng cách tập trung vào việc hiển thị các thông tin quan trọng, sử dụng các biểu đồ phù hợp và cung cấp các yếu tố tương tác, bạn có thể tạo ra các dashboard và report hiệu quả, hỗ trợ tốt cho việc ra quyết định dựa trên dữ liệu.

**Tối ưu hóa dashboard**

**Giao diện trực quan và hấp dẫn**

* **Sử dụng bố cục đơn giản:** Tránh quá nhiều chi tiết và yếu tố không cần thiết.
* **Sử dụng màu sắc hợp lý:** Chọn màu sắc tương phản tốt để làm nổi bật các thông tin quan trọng nhưng không gây rối mắt.

**Tập trung vào thông tin quan trọng**

* **Sử dụng biểu đồ phù hợp:** Chọn đúng loại biểu đồ cho từng loại dữ liệu (ví dụ: biểu đồ đường cho xu hướng, biểu đồ cột cho so sánh).

**Tương tác và bộ lọc**

* **Thêm các bộ lọc:** Cho phép người dùng lọc dữ liệu theo các tiêu chí khác nhau (ví dụ: theo thời gian, theo hãng xe).
* **Sử dụng các yếu tố tương tác:** Cho phép người dùng tương tác với các biểu đồ để xem chi tiết hơn (ví dụ: drill-down, hover để xem thêm thông tin).

**Hiệu suất**

* **Tối ưu hóa dữ liệu nguồn:** Chỉ lấy dữ liệu cần thiết để tránh làm chậm dashboard.
* **Sử dụng các biện pháp tối ưu hóa:** Như tính toán trước các measure phức tạp và lưu vào bộ nhớ đệm.

## xây dựng báo cáo

### dashboard và report

**Cách tối ưu hóa Dashboard:**

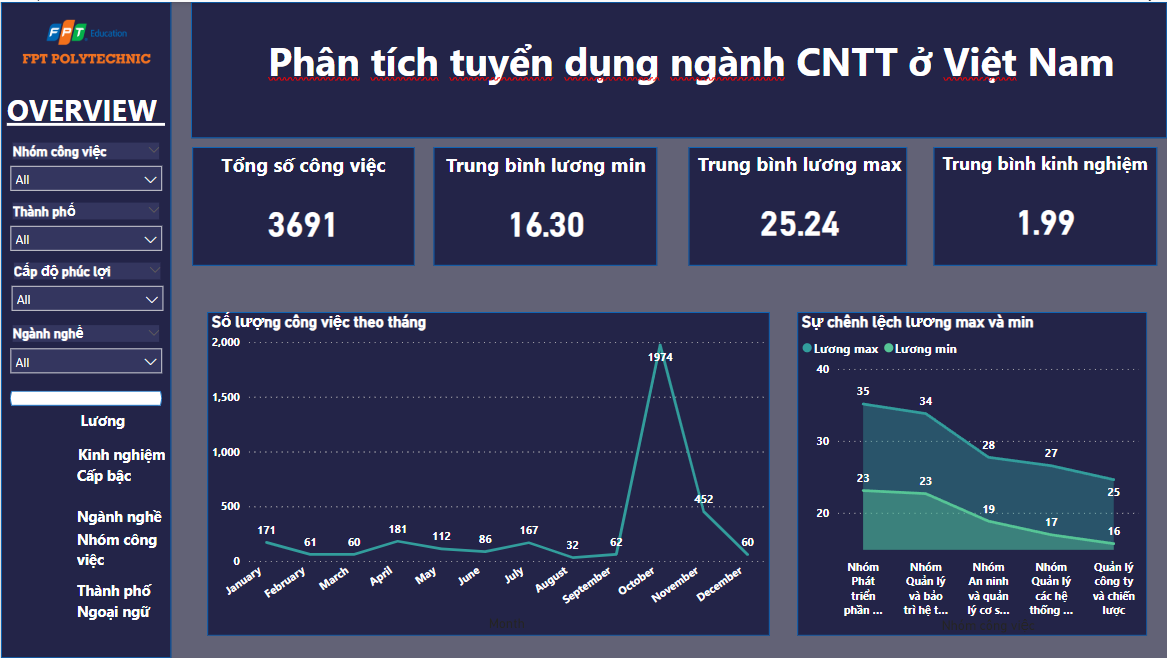
* **Sử dụng trực quan phù hợp:** Chọn biểu đồ và đồ thị phù hợp để truyền đạt thông tin một cách hiệu quả. Ví dụ, sử dụng biểu đồ đường để theo dõi xu hướng theo thời gian, biểu đồ cột để so sánh các danh mục và biểu đồ tròn để hiển thị tỷ lệ phần trăm.
* **Sắp xếp hợp lý:** Sắp xếp các biểu đồ và đồ thị một cách logic và dễ hiểu. Đặt các thông tin quan trọng nhất ở vị trí nổi bật và sử dụng tiêu đề, chú thích rõ ràng để giải thích nội dung.
* **Thiết kế đơn giản:** Tránh sử dụng quá nhiều chi tiết và đồ họa không cần thiết. Thiết kế dashboard đơn giản, dễ nhìn và tập trung vào thông tin quan trọng.

### dashboard

Trong dự án này, chúng tôi sẽ tạo ra các loại bảng điều khiển khác nhằm mục đích phục vụ các mục đích cụ thể và đáp ứng nhu cầu của các bên liên quan. Việc tạo ra các loại bảng điều khiển khác nhau giúp công ty Olist có cái nhìn toàn diện và chi tiết về các khía cạnh khác nhau của hoạt động kinh doanh. Mỗi bảng điều khiển phục vụ một mục tiêu cụ thể, từ tổng doanh thu, hiệu suất đơn hàng, hành vi khách hàng, hiệu quả nhà bán hàng đến phân tích tiếp thị. Điều này không chỉ giúp cải thiện hiệu quả quản lý và quyết định mà còn tăng cường khả năng cạnh tranh và phát triển sự bền vững của công ty.

Dưới đây là các loại dashboard:

#### Dashboard tổng quan (overview dashboard)



*Hình 6.1: Overview*

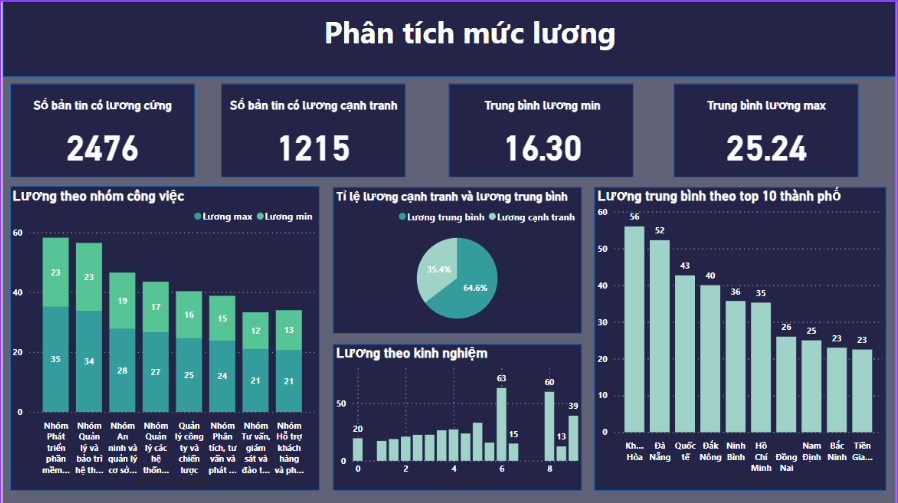
**Nội dung:**

* Biểu đồ slicing cho các trường: Nhóm công việc, Thành phố, Cấp độ phúc lợi
* Biểu đồ Card: thể hiện Tổng số công việc, Trung bình lương min, Trung bình lương max, Trung bình kinh nghiệm
* Biểu đồ đường tính số lượng công việc theo tháng
* Biểu đồ miền thể hiện sự chênh lệch lương max và lương min

**Mục đích:**

* Theo dõi xu hướng tuyển dụng ngành CNTT tại Việt Nam
* Hỗ trợ phân tích và so sánh
* Cung cấp thông tin chi tiết phục vụ ra quyết định
* Hỗ trợ định hướng nghề nghiệp và chiến lược tuyển dụng

#### Dashboard phân tích mức lương (Salary)

*Hình 6.2: Phân tích mức lương*

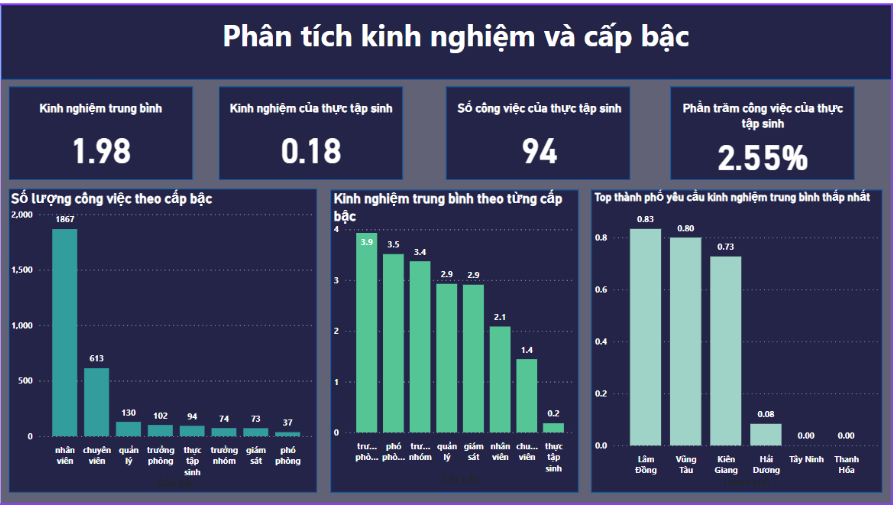
**Nội dung:**

* Biểu đồ cột hiển thị mức lương trung bình theo top 10 thành phố
* Biểu đồ card tính số bản tin có mức lương cứng, mức lương cạnh tranh, trung bình lương max và lương min
* Biểu đồ tròn tính tỷ lệ lương cạnh tranh và trung bình
* Bản đồ cột tính lương phân theo kinh nghiệm
* Biểu đồ cột chồng tính lương theo nhóm công việc

**Mục đích:**

* Hỗ trợ nhà tuyển dụng
* Hỗ trợ người tìm việc
* Hiểu rõ xu hướng thị trường
* Hỗ trợ chính sách phát triển ngành CNTT

#### Dashboard phân tích kinh nghiệm và cấp bậc

*Hình 6.3: Phân tích kinh nghiệm và cấp bậc*

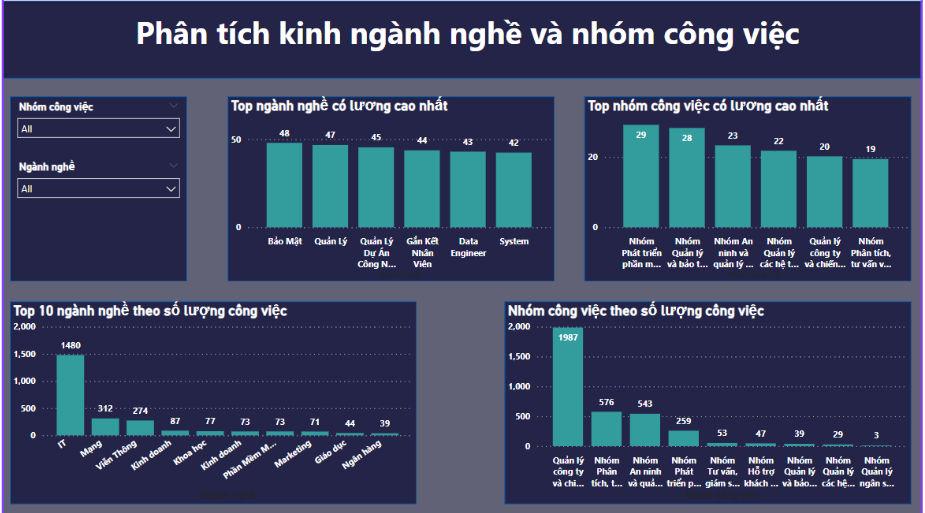
**Nội dung:**

* Biểu đồ Card thể hiện số năm kinh nghiệm, yêu cầu kinh nghiệm thực tập sinh ít nhất 3 tháng, số lượng công việc thực tập sinh và phần trăm công việc của thực tập sinh.
* Biểu đồ cột thể hiện top 6 thành phố yêu cầu kinh nghiệm trung bình thấp nhất.
* Biểu đồ cột thể hiện số lượng công việc theo cấp bậc
* Biểu đồ cột thể hiện kinh nghiệm trung bình theo cấp bậc

**Mục đích:**

* Hiểu rõ yêu cầu thị trường tuyển dụng
* Hỗ trợ người tìm việc
* Hỗ trợ nhà tuyển dụng
* Hỗ trợ phân tích xu hướng dài hạn
* Hỗ trợ chính sách đào tạo

#### Dashboard phân tích ngành nghề và nhóm công việc

*****Hình 6.4: Phân tích ngành nghề và nhóm công việc*

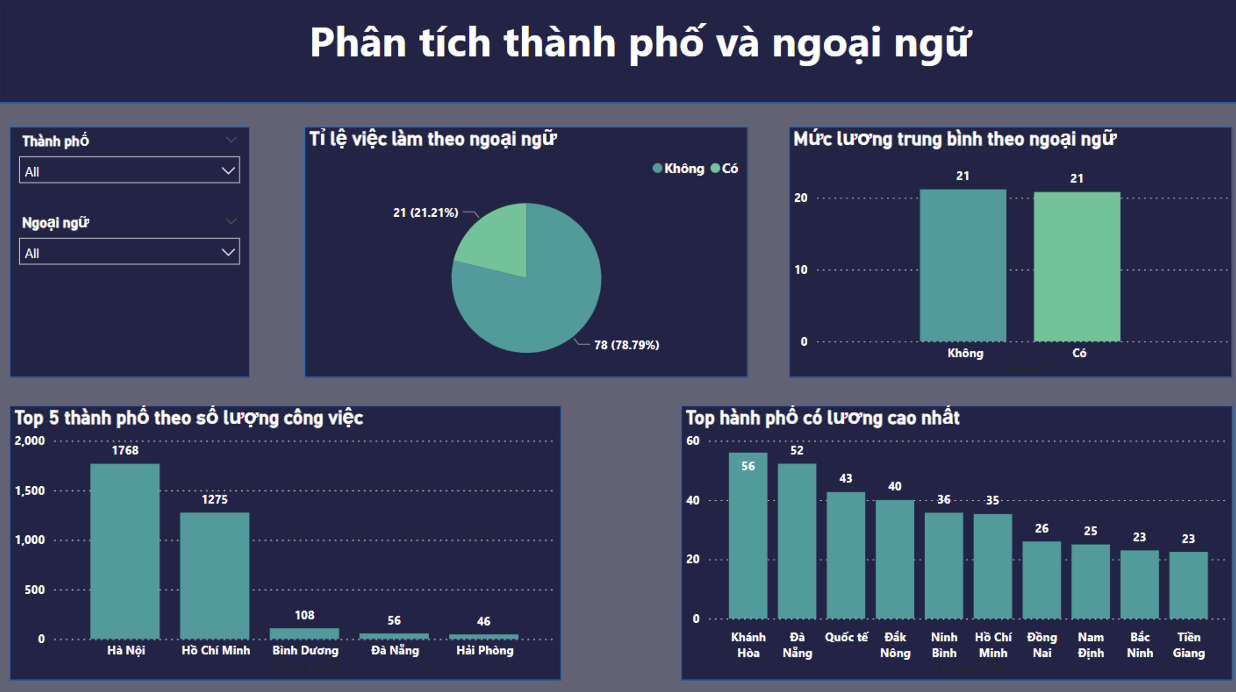
**Nội dung:**

* Biểu đồ cột slicer thể hiện nhóm công việc và ngành nghề.
* Biểu đồ cột thể hiện top 10 ngành nghề theo số trường công việc.
* Biểu đồ cột thể hiện nhóm công việc theo số lượng công việc.
* Biểu đồ cột thể hiện top 6 ngành nghề có số lương cao nhất.

**Mục đích:**

* Hiểu rõ nhu cầu tuyển dụng theo ngành nghề và công việc.
* Phân tích mức độ cạnh tranh trong ngành và công việc.
* Hỗ trợ người tìm việc.
* Hỗ trợ nhà tuyển dụng.
* Hỗ trợ phân tích xu hướng ngành nghề lâu dài.
* Hỗ trợ chính sách đào tạo và phát triển.

#### -Dashboard phân tích thành phố và ngoại ngữ

*Hình 6.5: Phân tích thành phố và ngoại ngữ*

**Nội dung:**

* Biểu đồ cột thể hiện top 5 thành phố theo công việc
* Biểu đồ cột thể hiện lương cao nhất theo thành phố
* Biểu đồ cột thể hiện mức lương trung bình theo thành phố
* Biểu đồ tròn hiển thị tỉ lệ việc làm theo ngoại ngữ

**Mục đích:**

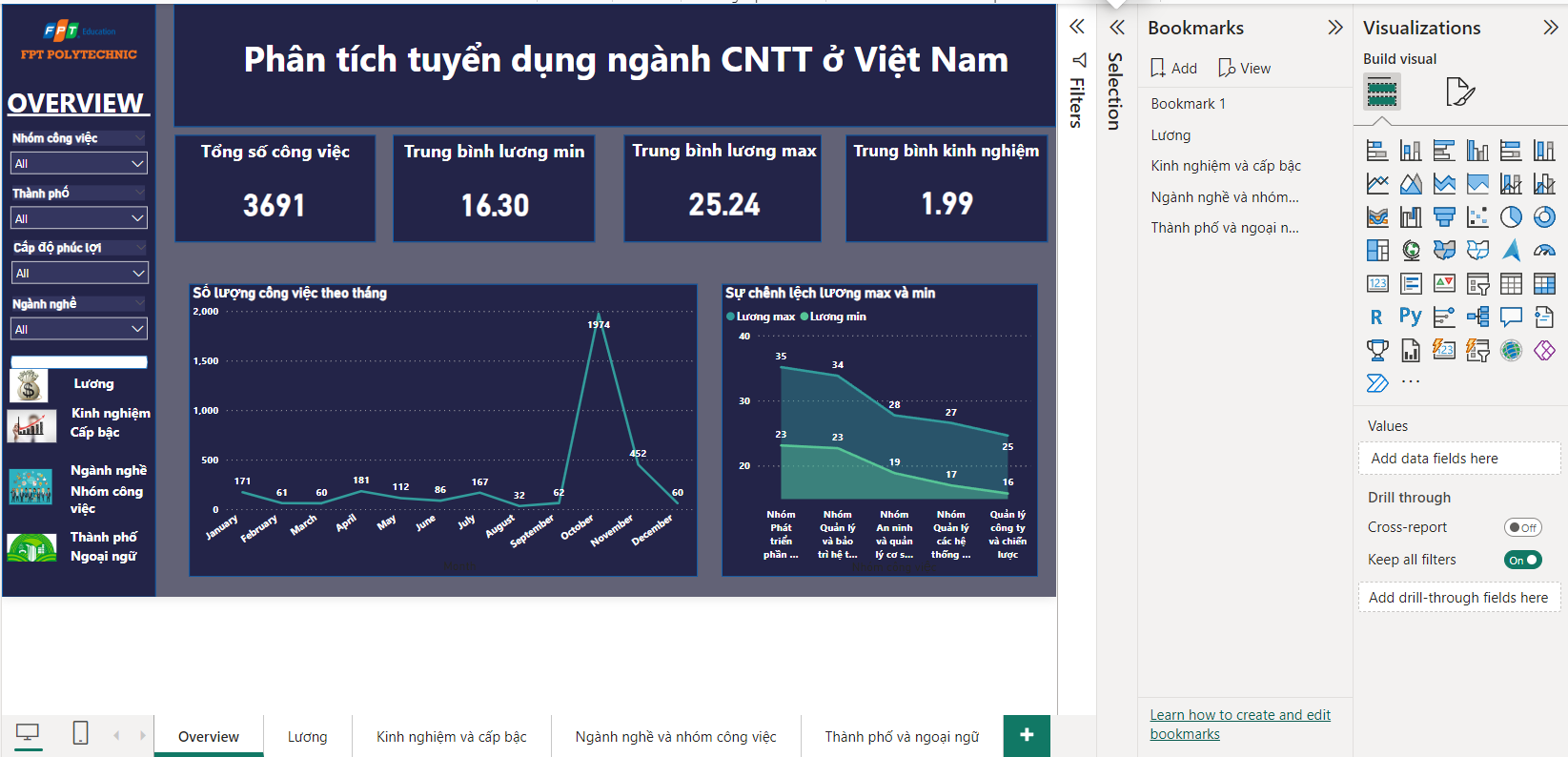
* Hiểu rõ yêu cầu thị trường tuyển dụng
* Hỗ trợ người tìm việc
* Hỗ trợ nhà tuyển dụng
* Hỗ trợ chính sách đào tạo

***Giải thích:***

Phân tích các trường thông tin như lương, kinh nghiệm, cấp bậc, nhóm ngành, và nhóm công việc giúp tạo ra bức tranh toàn diện về thị trường tuyển dụng. **Lương** là yếu tố quan trọng, phản ánh giá trị thị trường của một công việc hoặc ngành nghề. Việc phân tích mức lương (tối thiểu, tối đa, và trung bình) không chỉ giúp xác định sự chênh lệch thu nhập giữa các ngành hoặc nhóm công việc, mà còn hỗ trợ người lao động lựa chọn công việc phù hợp với mục tiêu tài chính. Đồng thời, doanh nghiệp có thể sử dụng thông tin này để xây dựng chiến lược lương bổng cạnh tranh, thu hút nhân tài.

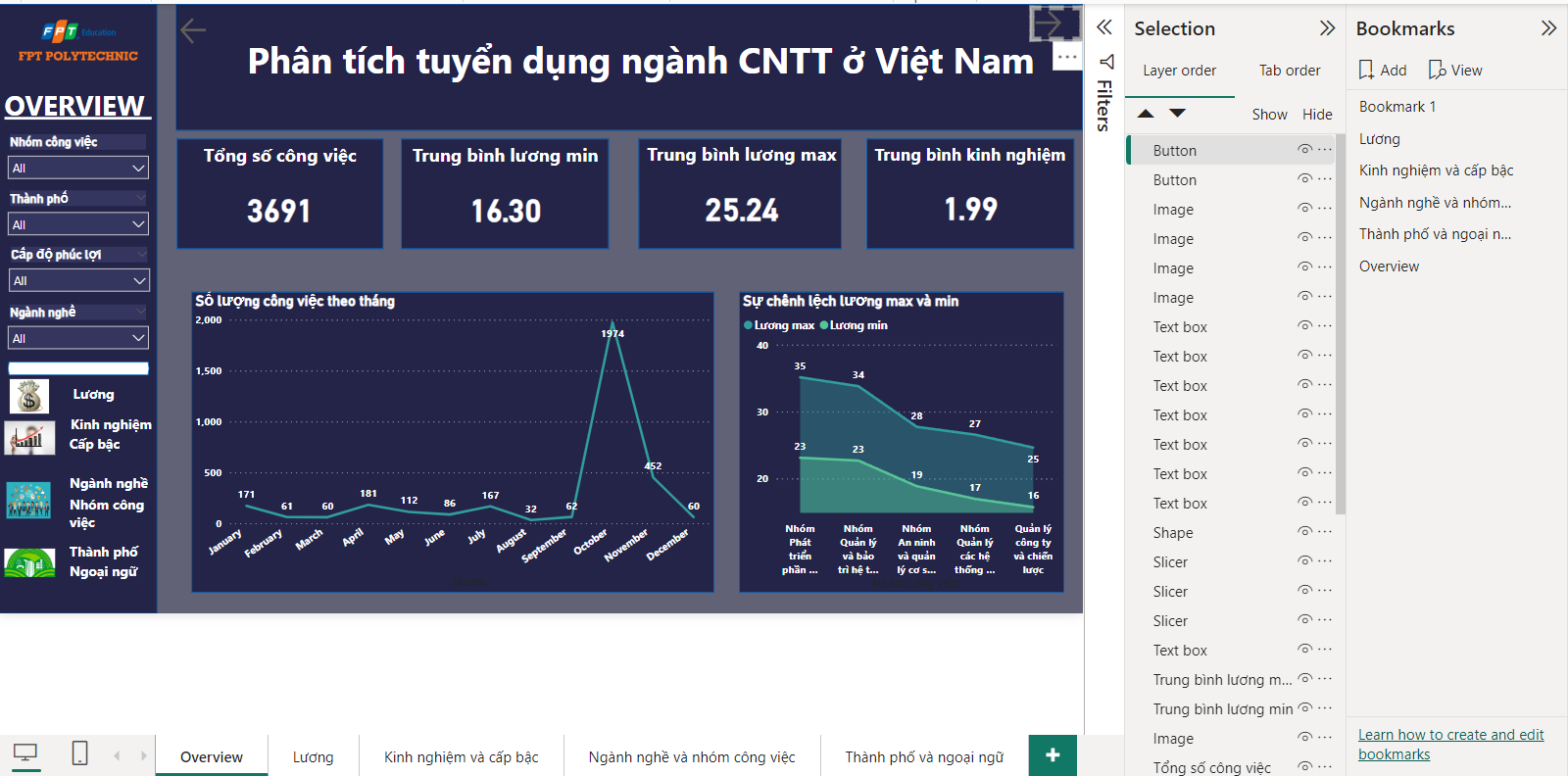
### bookmark

#### Tạo bookmark overview

*Hình 6.6: Bookmark overview*

*Giải thích:* Dùng để mở menu chứa các button dẫn đến các trang dashboardtrang.

#### Tạo bookmark chuyển tiếp

*Hình 6.7: Bookmark chuyển tiếp*

*Giải thích:* Dùng để đóng chuyển tiếp giữa các trang.

# kết luận

## báo cáo

### các bước viết báo cáo

Viết báo cáo phân tích dữ liệu đòi hỏi một quy trình có cấu trúc để đảm bảo rằng báo cáo được trình bày rõ ràng, mạch lạc và có giá trị đối với người đọc.Việc viết báo cáo phân tích dữ liệu đòi hỏi một quy trình có cấu trúc và cẩn thận. Bằng cách tuân theo các bước, bạn có thể tạo ra một báo cáo phân tích dữ liệu chất lượng, giúp truyền đạt các phát hiện quan trọng một cách rõ ràng và hiệu quả, từ đó hỗ trợ việc ra quyết định và cải thiện hiệu suất kinh doanh. Dưới đây là các bước cụ thể để viết một báo cáo phân tích dữ liệu:

**Bước 1: Xác định mục tiêu báo cáo**

**Xác định câu hỏi nghiên cứu:** Hiểu rõ mục tiêu của báo cáo và các câu hỏi chính cần trả lời.

**Đối tượng đọc báo cáo:** Xác định ai sẽ đọc báo cáo và điều chỉnh nội dung cho phù hợp với nhu cầu và mức độ hiểu biết của họ.

**Bước 2: Thu thập dữ liệu**

**Chọn nguồn dữ liệu:** Xác định các nguồn dữ liệu cần thiết để trả lời các câu hỏi nghiên cứu.

**Thu thập dữ liệu:** Thu thập dữ liệu từ các nguồn đã xác định.

**Bước 3: Chuẩn bị dữ liệu**

**Làm sạch dữ liệu:** Kiểm tra và xử lý các lỗi, giá trị thiếu, và giá trị ngoại lai trong dữ liệu.

**Chuẩn hóa dữ liệu:** Đảm bảo rằng dữ liệu được định dạng đúng cách và sẵn sàng cho phân tích.

**Bước 4: Phân tích dữ liệu**

**Lựa chọn phương pháp phân tích:** Chọn các phương pháp phân tích phù hợp với mục tiêu báo cáo (ví dụ: phân tích mô tả, phân tích hồi quy, phân tích xu hướng).

**Thực hiện phân tích:** Sử dụng các công cụ và kỹ thuật phù hợp để phân tích dữ liệu.

**Bước 5: Trực quan hóa dữ liệu**

**Chọn loại biểu đồ phù hợp:** Chọn các loại biểu đồ và đồ thị phù hợp để trực quan hóa dữ liệu (ví dụ: biểu đồ đường, biểu đồ cột, biểu đồ tròn).

**Tạo biểu đồ:** Sử dụng các công cụ trực quan hóa dữ liệu để tạo biểu đồ minh họa cho các phát hiện quan trọng.

**Bước 6: Viết báo cáo**

**Giới thiệu:**

* Trình bày mục tiêu của báo cáo.
* Giới thiệu ngắn gọn về dữ liệu và phương pháp phân tích.

**Phương pháp:** Mô tả chi tiết các phương pháp và công cụ sử dụng trong quá trình phân tích.

**Kết quả:**

* Trình bày các phát hiện chính từ phân tích dữ liệu.
* Sử dụng biểu đồ và đồ thị để minh họa cho các phát hiện.

**Thảo luận:**

* Giải thích ý nghĩa của các phát hiện.
* Đề xuất các hành động hoặc quyết định dựa trên kết quả phân tích.

**Kết luận:**

* Tóm tắt lại các điểm chính của báo cáo.
* Nêu rõ các kết luận chính và các bước tiếp theo nếu có.

**Bước 7: Xem xét và chỉnh sửa**

**Đọc lại báo cáo:** Kiểm tra lại báo cáo để đảm bảo rằng nội dung rõ ràng, chính xác và không có lỗi.

**Chỉnh sửa:** Sửa các lỗi ngữ pháp, chính tả và đảm bảo rằng báo cáo có cấu trúc logic.

**Bước 8: Trình bày báo cáo**

**Định dạng:** Đảm bảo rằng báo cáo được định dạng chuyên nghiệp và dễ đọc.

**Trình bày:** Chuẩn bị sẵn sàng để trình bày báo cáo trước các bên liên quan nếu cần thiết.

### tổng hợp

Tổng hợp phân tích dữ liệu tuyển dụng ngành công nghệ thông tin tại Việt Nam:

* Phân tích dữ liệu cho thấy các yếu tố quan trọng ảnh hưởng đến việc tuyển dụng và hiệu quả của ngành công nghệ thông tin. Thời gian đăng tin tuyển dụng và mức độ cập nhật đóng vai trò quan trọng, khi các vị trí được đăng gần đây thường có mức lương cạnh tranh hơn, phản ánh sự tăng trưởng của ngành. Các nhóm công việc như Phát triển phần mềm, Quản lý dự án, và Kỹ thuật hệ thống có nhu cầu cao hơn và mức lương hấp dẫn hơn so với các nhóm công việc như Hỗ trợ kỹ thuật hoặc Nhập liệu. Các ngành nghề công nghệ cao, đặc biệt là IT - Phần mềm, AI, và Phân tích dữ liệu, mang lại giá trị tuyển dụng vượt trội so với các ngành nghề như kinh doanh hoặc tiếp thị trong lĩnh vực công nghệ.
* Yếu tố địa lý cũng ảnh hưởng đáng kể, khi các công ty tập trung tại Hà Nội và TP.HCM chiếm phần lớn cơ hội tuyển dụng với mức lương và phúc lợi hấp dẫn hơn các khu vực khác. Phân tích cũng chỉ ra rằng các phúc lợi như bảo hiểm, du lịch, và các khoản phụ cấp giúp tăng sức hút tuyển dụng, đặc biệt với các ứng viên có kinh nghiệm từ 2 đến 3 năm.
* Mức độ hài lòng và yêu cầu công việc được đánh giá qua cấp bậc và mức lương cho thấy các công việc yêu cầu kinh nghiệm cao hơn thường cung cấp mức lương tốt hơn. Tuy nhiên, các vị trí yêu cầu ngoại ngữ vẫn chưa phổ biến, dẫn đến tiềm năng hạn chế cho việc mở rộng thị trường quốc tế.

Các phân tích cũng chỉ ra rằng:

* **Nhóm công việc ảnh hưởng đến kết quả xử lý hồ sơ**: Các nhóm công việc như **Phát triển phần mềm**, **Quản lý dự án**, và **Kỹ thuật hệ thống** thường yêu cầu thời gian xử lý hồ sơ nhanh chóng để không mất đi nhân tài, do đây là các lĩnh vực có sự cạnh tranh nhân lực rất cao.
* **Cấp bậc và yêu cầu kinh nghiệm ảnh hưởng đến khả năng tuyển dụng**: Các vị trí ở cấp quản lý hoặc yêu cầu kinh nghiệm cao thường khó tuyển dụng hơn, mặc dù mức lương hấp dẫn.
* **Yếu tố ngoại ngữ là rào cản trong nhiều vị trí tuyển dụng**: Dù không phải yêu cầu phổ biến, nhưng các vị trí cần ngoại ngữ, đặc biệt là tiếng Anh, thường hạn chế số lượng ứng viên đủ điều kiện, dẫn đến thời gian tuyển dụng kéo dài hơn.
* **Phân bổ địa lý chưa đồng đều**: Các cơ hội tuyển dụng tập trung mạnh tại Hà Nội và TP.HCM, trong khi các khu vực khác vẫn thiếu cơ hội, tạo ra sự cạnh tranh gay gắt ở hai thành phố lớn.

Những phát hiện này nhấn mạnh nhu cầu tối ưu hóa quy trình tuyển dụng, cải thiện chính sách đãi ngộ và xây dựng chiến lược dài hạn để tiếp cận nguồn nhân lực từ các khu vực khác nhau.

## khó khăn

Việc tuyển dụng trong ngành CNTT cũng đối mặt với nhiều khó khăn, bao gồm yêu cầu cao về kỹ năng, kinh nghiệm và đôi khi là ngoại ngữ, làm hạn chế cơ hội cho những ứng viên trẻ hoặc mới ra trường. Mức lương không đồng đều giữa các vị trí, trong đó một số công việc có mức lương thấp hơn kỳ vọng, làm giảm sức hút đối với nhân lực tài năng. Ngoài ra, sự tập trung tuyển dụng tại các thành phố lớn khiến ứng viên ở các khu vực xa trung tâm gặp khó khăn trong việc tiếp cận cơ hội. Hơn nữa, sự cạnh tranh giữa các doanh nghiệp để thu hút nhân tài cũng làm gia tăng áp lực, trong khi một số công việc bị phân mảnh hoặc không rõ ràng về chuyên môn ngành CNTT, gây nhầm lẫn và khó khăn trong việc tuyển đúng người phù hợp.

## thuận lợi

Ngành CNTT tại Việt Nam hiện có nhiều thuận lợi trong tuyển dụng, bao gồm sự đa dạng trong các vị trí công việc từ cấp nhân viên đến quản lý cao cấp, giúp đáp ứng được nhu cầu của nhiều nhóm ứng viên. Các công việc thường đi kèm với mức lương cạnh tranh, thậm chí rất cao ở một số vai trò quan trọng, cùng với các phúc lợi hấp dẫn như bảo hiểm, du lịch, phụ cấp, và trang bị thiết bị làm việc. Tập trung tuyển dụng chủ yếu tại các thành phố lớn như Hà Nội và TP.HCM giúp các công ty tiếp cận với nguồn nhân lực chất lượng cao. Điều này cũng phản ánh sự phát triển mạnh mẽ của thị trường lao động CNTT tại Việt Nam.

## hướng phát triển

Dựa trên kết quả phân tích dữ liệu tuyển dụng ngành công nghệ thông tin tại Việt Nam, dưới đây là một số hướng phát triển cho dự án phân tích dữ liệu, bên cạnh việc tập trung vào sự phát triển của ngành công nghệ thông tin:

#### Mở rộng phạm vi phân tích:

* **Phân tích chi tiết hơn về hành vi tuyển dụng**: Phân khúc ứng viên theo ngành nghề, xây dựng mô hình dự đoán xu hướng tuyển dụng trong tương lai, dự báo nhu cầu nhân sự trong các lĩnh vực công nghệ cao như AI, Big Data, Phát triển phần mềm.
* **Phân tích hiệu quả tuyển dụng của từng công ty**: Đánh giá mức độ thành công trong việc tuyển dụng, phân tích các vị trí tuyển dụng có mức độ cạnh tranh cao, và đưa ra các đề xuất cải thiện quy trình tuyển dụng cho các công ty.
* **Phân tích chiến lược tuyển dụng**: Đánh giá hiệu quả của các chiến dịch tuyển dụng qua các kênh như website tuyển dụng, mạng xã hội, hoặc qua các sự kiện tuyển dụng. Tối ưu hóa chi phí tuyển dụng và xác định kênh quảng cáo hiệu quả nhất.
* **Phân tích dữ liệu theo khu vực địa lý**: Phân tích nhu cầu tuyển dụng theo các khu vực (Hà Nội, TP.HCM, các tỉnh khác), tối ưu hóa chiến lược tuyển dụng dựa trên xu hướng nhân sự khu vực.

#### Nâng cao chất lượng dữ liệu:

* **Thu thập thêm dữ liệu**: Bổ sung thông tin về kỹ năng chuyên môn của ứng viên, nền tảng giáo dục, chứng chỉ và các thông tin bổ sung như sự tham gia vào các dự án công nghệ.
* **Cải thiện chất lượng dữ liệu**: Xử lý dữ liệu bị thiếu, xác định và loại bỏ dữ liệu ngoại lai, đảm bảo tính nhất quán và độ chính xác của thông tin ứng viên.
* **Tự động hóa quy trình làm sạch và chuẩn hóa dữ liệu**: Tăng cường hiệu quả xử lý dữ liệu bằng các công cụ tự động để đảm bảo chất lượng dữ liệu luôn ở mức cao.

#### Phát triển mô hình dự đoán:

* **Dự đoán nhu cầu tuyển dụng**: Xây dựng mô hình dự đoán nhu cầu nhân sự theo ngành nghề trong những năm tới, hỗ trợ các công ty trong việc lập kế hoạch nhân sự và chiến lược tuyển dụng.
* **Dự đoán xu hướng phát triển nghề nghiệp**: Hỗ trợ ứng viên dự báo các kỹ năng và công việc sẽ được yêu cầu trong tương lai, từ đó giúp họ lên kế hoạch phát triển nghề nghiệp.

#### Cải thiện trực quan hóa dữ liệu:

* **Sử dụng nhiều kỹ thuật trực quan hóa khác nhau**: Tạo ra các bảng điều khiển tương tác (dashboards), sử dụng biểu đồ và bản đồ phân tích dữ liệu nhân sự để đưa ra các thông tin dễ hiểu về tình hình tuyển dụng và thị trường lao động.
* **Thiết kế giao diện báo cáo thân thiện với người dùng**: Tạo giao diện trực quan, dễ dàng tìm kiếm và phân tích thông tin về ngành nghề, nhóm công việc, và mức độ tuyển dụng tại các công ty.
* **Phát triển ứng dụng trực quan hóa dữ liệu trên thiết bị di động**: Cung cấp báo cáo và dữ liệu phân tích dễ dàng truy cập thông qua các thiết bị di động cho các nhà tuyển dụng và ứng viên.

#### Nâng cao trải nghiệm người tìm việc:

* **Tối ưu hóa quy trình ứng tuyển**: Cải thiện thời gian phản hồi và xử lý hồ sơ ứng viên, hợp tác với các công ty tuyển dụng để tăng cường hiệu quả ứng tuyển.
* **Cải thiện dịch vụ hỗ trợ ứng viên**: Đầu tư vào hệ thống hỗ trợ tư vấn nghề nghiệp cho ứng viên, cung cấp các thông tin về các cơ hội nghề nghiệp và giúp ứng viên phát triển kỹ năng.
* **Cá nhân hóa trải nghiệm tuyển dụng**: Phân tích hành vi và sở thích của ứng viên để đưa ra các công việc phù hợp, cá nhân hóa các cơ hội nghề nghiệp và tạo ra một quy trình ứng tuyển đơn giản và nhanh chóng.

#### Tối ưu hóa hoạt động tuyển dụng:

* **Tập trung vào các kênh tuyển dụng hiệu quả**: Phân tích hiệu quả của các kênh tuyển dụng, đầu tư vào các nền tảng tuyển dụng mang lại tỷ lệ ứng viên chất lượng cao.
* **Phát triển các chiến dịch tuyển dụng sáng tạo**: Thu hút ứng viên mới, kích thích nhu cầu ứng tuyển, và nâng cao nhận diện thương hiệu của công ty trong ngành công nghệ.
* **Xây dựng chương trình khuyến khích ứng viên**: Khuyến khích ứng viên tham gia vào các chương trình đào tạo, phát triển nghề nghiệp, hoặc các hoạt động gắn kết lâu dài với công ty.

#### Mở rộng thị trường và danh mục ngành nghề:

* **Mở rộng sang các khu vực tuyển dụng mới**: Phân tích thị trường lao động tại các tỉnh thành chưa được khai thác mạnh, phát triển các cơ hội tuyển dụng tại các khu vực ngoài Hà Nội và TP.HCM.
* **Mở rộng danh mục ngành nghề**: Cung cấp các cơ hội nghề nghiệp đa dạng cho các ứng viên, mở rộng ra các ngành nghề khác ngoài công nghệ như marketing kỹ thuật số, quản lý sản phẩm, v.v.
* **Hợp tác với các tổ chức đào tạo**: Mở rộng hợp tác với các trường đại học, cơ sở đào tạo nghề để cung cấp nguồn nhân lực chất lượng cao cho thị trường.

#### Nâng cao hiệu quả hoạt động tuyển dụng:

* **Tối ưu hóa quy trình tuyển dụng**: Sử dụng công nghệ để tự động hóa các quy trình tuyển dụng, từ việc lọc hồ sơ đến phỏng vấn, giảm thiểu lỗi và tăng cường hiệu quả.
* **Quản lý dữ liệu ứng viên hiệu quả**: Dự báo nhu cầu nhân sự theo từng ngành, tối ưu hóa việc tuyển dụng và lưu trữ dữ liệu ứng viên.
* **Phân tích dữ liệu tuyển dụng thường xuyên**: Theo dõi các chỉ số hiệu quả tuyển dụng, nhận diện các vấn đề và cơ hội, điều chỉnh chiến lược tuyển dụng kịp thời.

Bằng cách tập trung vào các hướng phát triển này, ngành công nghệ thông tin tại Việt Nam có thể cải thiện hiệu quả tuyển dụng, đáp ứng nhu cầu nhân lực ngày càng cao, và tăng cường sự phát triển bền vững trong lĩnh vực này.

# tổng kết

**Dự án: Phân tích xu hướng và hành vi người tiêu dùng trên trang thương mại điện tử Olist tại Brazil**

Olist, nền tảng thương mại điện tử hàng đầu tại Brazil, đã thực hiện một dự án phân tích dữ liệu sâu rộng nhằm khai thác tối đa tiềm năng của bộ dữ liệu Brazilian E-Commerce Public Dataset và Marketing Funnel. Mục tiêu chính của dự án là:

* **Hiểu rõ hành vi khách hàng:** Phân tích chi tiết hành vi mua sắm, sở thích và thói quen của khách hàng để xây dựng các chiến lược tiếp thị hiệu quả.
* **Tối ưu hóa quy trình bán hàng:** Nhận diện các điểm hạn chế trong quy trình bán hàng và đề xuất các giải pháp cải tiến để tăng tỷ lệ chuyển đổi.
* **Cải thiện hiệu quả tiếp thị:** Đánh giá hiệu quả của các chiến dịch marketing và tối ưu hóa ngân sách quảng cáo.

**Phương pháp thực hiện**

Dự án đã áp dụng một quy trình phân tích dữ liệu toàn diện, bao gồm:

1. **Thu thập và làm sạch dữ liệu:** Sưu tập dữ liệu từ nhiều nguồn khác nhau và tiến hành làm sạch, chuẩn hóa để đảm bảo tính chính xác và nhất quán.
2. **Mô hình hóa dữ liệu:** Xây dựng các mô hình thống kê và máy học để phát hiện các mối quan hệ phức tạp giữa các yếu tố ảnh hưởng đến doanh số.
3. **Trực quan hóa dữ liệu:** Sử dụng Power BI để tạo ra các báo cáo trực quan, dễ hiểu, giúp người dùng nắm bắt nhanh chóng các thông tin quan trọng.

**Kết quả đạt được**

Qua quá trình phân tích, dự án đã đạt được những kết quả đáng kể:

* **Xác định các yếu tố ảnh hưởng đến doanh số:** Thời gian mua hàng, danh mục sản phẩm, vị trí địa lý, phương thức thanh toán và mức độ hài lòng của khách hàng đều có tác động đáng kể đến giá trị đơn hàng.
* **Đánh giá hiệu quả các chương trình khuyến mãi:** Các chương trình khuyến mãi đã được chứng minh là có hiệu quả trong việc thúc đẩy doanh số và cải thiện trải nghiệm khách hàng.
* **Nhận diện các điểm hạn chế trong quy trình bán hàng:** Sự chậm trễ trong việc xử lý đơn hàng quá lâu hay dịch vụ giao hàng kéo dài là một trong những nguyên nhân khiến khách hàng không hài lòng và rời bỏ.

**Các gợi ý phát triển**

Dựa trên kết quả phân tích, Olist có thể tập trung vào các hướng phát triển sau:

* **Mở rộng phạm vi phân tích:** Áp dụng các kỹ thuật phân tích tiên tiến hơn để khám phá sâu hơn vào dữ liệu.
* **Nâng cao chất lượng dữ liệu:** Đầu tư vào việc thu thập và quản lý dữ liệu một cách hiệu quả.
* **Phát triển mô hình dự đoán:** Xây dựng các mô hình dự đoán để dự báo doanh số, hành vi khách hàng và xu hướng thị trường.
* **Cải thiện trực quan hóa dữ liệu:** Tạo ra các dashboard tương tác và trực quan hơn để hỗ trợ ra quyết định.
* **Nâng cao trải nghiệm khách hàng:** Tối ưu hóa quy trình bán hàng, dịch vụ khách hàng và các chương trình khuyến mãi.
* **Tối ưu hóa hoạt động tiếp thị:** Tập trung vào các kênh tiếp thị hiệu quả và cá nhân hóa trải nghiệm khách hàng.
* **Mở rộng thị trường và danh mục sản phẩm:** Xâm nhập vào các thị trường mới và đa dạng hóa sản phẩm.
* **Nâng cao hiệu quả hoạt động:** Tối ưu hóa các quy trình nội bộ và giảm chi phí.

**Slice Dữ liệu thô Báo cáo**

[**III. Thiết kế CSDL 7**](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.vwx8bk9lp067)

Bảng1.1: Fact Table [7](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.46bu5hl26i4x)

Bảng1.2: Dimension Tables [9](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.1lln2uqwrfvc)

Bảng2.1: Fact\_Air\_Quality [12](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.bofr8tnxzkwx)

Bảng2.2: dim\_time [12](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.bofr8tnxzkwx)

Bảng2.3: Dim\_Location [12](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.bofr8tnxzkwx)

Bảng2.4: Dim\_Parameter [12](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.bofr8tnxzkwx)

Bảng2.5: Dim\_Provider [12](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.bofr8tnxzkwx)

[**IV. Xây dựng kho dữ liệu 12**](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.of26js1zlqks)

Bảng [1. Chọn các chiều dữ liệu (Dimension) 12](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.uwa6i7sg15tk)

Bảng [2. Xác định bảng sự kiện (Fact) 12](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.7j66mutq61d1)

Bảng [3. Tích hợp dữ liệu và cài đặt kho dữ liệu 13](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.jy4eakpau6jr)

Bảng [3.1. Tích hợp dữ liệu vào kho dữ liệu 13](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.m3ajebvzq647)

Bảng [3.2. Cài đặt kho dữ liệu 13](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.n1xj3gy6v48b)

[**V. Khai thác kho dữ liệu 13**](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.4rbcuopkmhfq)

[**VI. Tổng kết 13**](https://docs.google.com/document/d/1cC7fjTfj4GVcyedylsXUUdDST9TMRHGV/r/edit/edit#heading=h.gixxcphrtcnj)